RESEARCH ON SPEECH PERCEPTION

Technical Report No. 3

December 18, 1980

Department of Psychology

Indiana University

Bloomington, Indiana 47405

Supported by:

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

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

and

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

REPRESENTATIONS OF SPECTRAL CHANGE AS CUES TO PLACE OF ARTICULATION IN STOP CONSONANTS

by

DIANE KEWLEY-PORT

A dissertation submitted to the Graduate Faculty in Speech and Hearing Sciences in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1981

Abstract

REPRESENTATIONS OF SPECTRAL CHANGE
AS CUES TO PLACE OF ARTICULATION IN
STOP CONSONANTS

by

Diane Kewley-Port

Adviser: Professor Michael Studdert-Kennedy

The specification of the acoustic cues to place of articulation in stop consonants has continued to be a prominent issue in research on speech perception and production. Investigators have examined several different acoustic cues in the burst and formant transitions of initial stops. Burst cues are generally described from single spectral sections, whereas formant transition cues are characterized by frequency changes over time. Recently Stevens and Blumstein (1978; 1980) have claimed that invariant cues for place of articulation can be found in the initial portion of a CV syllable integrated over burst and transition segments. They argue that this invariant information can be specified by the gross shape of the spectrum at onset.

The premise underlying the present investigation is that change in the spectral distribution of energy over time is the proper characterization of the underlying articulatory gestures of stop consonants as well as the

neural representation of speech signals in the peripheral auditory system. Four experiments were designed to investigate acoustic cues characterized by spectral change and to compare these time-varying cues to the static onset spectra cues proposed by Stevens and Blumstein (1978).

The first experiment reexamined parameters of formant transitions as possible acoustic correlates to place of articulation. The purpose of this study was to use digital signal processing techniques to measure formant transitions in great detail from naturally produced initial stop-vowel syllables. Analysis of several formant transition parameters showed that they did not distinguish place of articulation across a number of vowel contexts.

The second experiment examined visual displays of the continuous change in spectral energy from the release burst into the formant transitions. Following a suggestion of Searle, Jacobson and Rayment (1979), three-dimensional running spectral displays were computer generated. These running spectra displayed 20 ms spectral sections of the waveforms offset in time at 5 ms intervals. A set of three time-varying features was defined. These features were shown to specify place of articulation accurately in initial stops over several vowel contexts and talkers. This analysis was compatible with other approaches to modeling the neural representation of speech signals that incorporate psychophysical properties of the human auditory system.

The last two experiments were designed to test several claims made by Stevens and Blumstein and to compare experimentally their proposal of static onset spectral cues for place with the dynamic spectral features from Experiment 2. In Experiment 3, observers identified either the consonant or the vowel in short initial portions of naturally produced stop-vowel syllables. The results demonstrated that sufficient acoustic information for identifying place of articulation is present in the first 20 to 40 ms of a stop waveform.

The final experiment compared consonant identification in two sets of synthetically produced CV stimuli. One set of stimuli preserved only the static onset spectra information as defined by Stevens and Blumstein. The other set was modeled after the time-varying features proposed in Experiment 2. The results showed that the stimuli synthesized from onset spectra alone did not contain reliable and sufficient acoustic cues to identify place of articulation. The time-varying stimuli were shown to contain salient perceptual information for identification of place of articulation.

Taken together, these results demonstrate that a set of dynamically changing temporal properties observable in initial portions of stop-vowel syllables can serve as reliable acoustic cues to place of articulation. These findings suggest that the appropriate descriptions of cues that listeners use to identify place of articulation should

incorporate dynamically changing spectral properties. The present results have implications for current models of speech perception and for descriptions of the neural filtering of speech signals by the peripheral auditory system.

ACKNOWLEDGMENTS

From the inception of this research, David B. Pisoni has made substantial personal and scientific commitments to the design and execution of this work. My debt of gratitude for his efforts is very great indeed. I owe him special thanks for providing vision and direction for this research, and I appreciate his efforts to fully support the technical aspects of this project. Finally, I am indebted for the careful attention he has devoted to the manuscript itself.

I am very grateful for the contributions Michael Studdert-Kennedy has made to this thesis. In particular, Experiment 3 was developed under his special guidance. I also greatly appreciate his careful reading of the final manuscript and his insightful comments in relating this work to theories of speech perception.

A special feeling of appreciation is given to Katherine Safford Harris. I thank her for many years of support, both as my friend and adviser. Her comments in many phases of the research and on the manuscript were probing and constructive.

In addition I thank Dennis H. Klatt and Lawrence J. Raphael for numerous contributions throughout the research, and for their thoughtful comments on the manuscript.

I am especially indebted to the support staff of the Speech Perception Laboratory, Department of Psychology,

Indiana University. Jerry C. Forshee developed the computer and other technical facilities. I also appreciate his assistance in the development of many of the research programs used in this thesis. I am especially grateful to Nancy Layman who not only assisted in running the laboratory, but was responsible for careful text processing of the manuscript. Thanks also to David Link for help with engineering and maintenance of the laboratory equipment, and to Suzanne Hull for careful preparation of the figures.

Finally I am most grateful to my husband Robert Port for his encouragement and wholehearted support while finishing this degree. He contributed directly to this research through discussions as a colleague in linguistics, and indirectly in ways too numerous to mention as a loving husband and father. I also wish to thank my children Nicholas, Juliet and Cynthia for their great patience and for the encouragement they have given me throughout this project.

The research reported in this dissertation was supported by from the National Institutes of Health, Grant Numbers NS 12179-05 and MH 24027-06.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

The specification of the acoustic cues to place of articulation in stop consonants has been a goal of research in speech perception and production for over thirty years. Despite extensive efforts, a satisfactory account of the salient perceptual cues for place of articulation has not emerged and this topic is still the subject of contemporary research in experimental phonetics. The description of the acoustic cues to place of articulation in the past has been based on two sources of information. In one approach, the acoustic properties have been calculated from theoretical models of the stop consonant articulatory gestures (Fant, 1960; Stevens and Blumstein, 1978). In the other, acoustic correlates have been measured from sound spectrograms or estimated using other analysis techniques (Cooper, Delattre, Liberman, Borst and Gerstman, 1952; Liberman, Delattre, Cooper and Gerstman, 1954). Although a great deal of detailed knowledge has accumulated over the years, it has been difficult to verify particular sets of acoustic properties as both necessary and sufficient perceptual cues to place of articulation for the human listener.

The research reported in this thesis is most closely related to recent attempts by Stevens and Blumstein (1978; 1980) and by Searle and his colleagues (Searle, Jacobson and Rayment, 1979; Searle, Jacobson and Kimberly, 1980) to resolve the problems of acoustic invariance in stop

consonants. Before we describe their research efforts, however, it will be useful to review briefly two basic premises regarding the nature of the stop-vowel waveform which contains the distinctive information required for the identification of place of articulation. Since the earliest spectrographic studies of Potter, Kopp and Green (1947) and Joos (1948), the acoustic cues for place of articulation in stop consonants have been thought to lie in two readily observable acoustic segments, the release burst and the formant transitions. These distinctive segments could be seen clearly in spectrographic displays of CV syllables. Cooper et al. (1952), Liberman et al.(1954), Liberman, Cooper, Shankweiler and Studdert-Kennedy, (1967), Cole and Scott (1974a) and many other investigators have made the assumption that these two segments are independent and separable. On the other hand, Fant (1960) and Stevens and Blumstein (1978) have assumed that these acoustic segments are not separable at all but constitute a unitary or integrated acoustic stimulus for specifying place in stops.

The premise that the release burst and formant transitions are separable derives primarily from the apparent distinctiveness of their visual representations in oscillograms and sound spectrograms. Typically, the spectral properties of the burst have been analyzed and measured from a single spectral section (Halle, Hughes, and Radley, 1957; Zue, 1976). Formant transitions have been

examined and measured primarily from sound spectrograms and perceptual cues have been described in terms of change in frequency over time (Potter et al., 1947; Joos, 1948; Liberman et al., 1954; Fant, 1973). Numerous speech perception studies have been carried out over the years to verify the role of these acoustic properties as cues to place. These studies have employed both synthetic and natural speech stimuli to determine the acoustic correlates that listeners find perceptually distinctive. The results of studies investigating formant transitions in isolation (Liberman et al., 1954) or with bursts (Cooper et al., 1952; Dorman, Studdert-Kennedy and Raphael, 1977) have shown that the acoustic information for place of articulation varies with the following vowel context. In contrast, other studies have focused on the burst in a limited number of vowel contexts (Cole and Scott, 1974a) and have concluded that most of the place information is located in the release burst. The more general conclusion, however, has been that both the burst and the formant transitions contribute to specifying place information in a complementary or "cue-trading" way depending on the following vowel context (see Dorman et al., 1977 or Fisher-Jørgensen, 1972).

The assumption that the burst and formant transitions constitute a single integrated acoustic stimulus has motivated a set of different experiments and has led to the development of several new theories of segmental speech

perception. Both Fant (1960; 1973) and Stevens and Blumstein (1978; Blumstein and Stevens, 1979) have argued that the acoustic information for specifying place of articulation is independent of vowel context and is located in the first 10 to 30 ms of the stop consonant waveform. Their claims are derived, in part, from theoretical predictions developed from the acoustic theory of speech production (see Fant, 1960; Stevens and Blumstein, 1978). Recent experimental findings by Stevens and Blumstein have been interpreted as support for this view, a view that argues for the existence of invariant cues for the perception of place in stop consonants.

Three recent studies formed the background for the research described in this thesis. The first was a perceptual experiment conducted by Dorman, Studdert-Kennedy and Raphael (1977) which used natural speech stimuli. In an extension of earlier studies, these investigators employed a large data base using the consonants /bdg/ before nine vowels spoken by two male talkers. The experiment consisted of excising and recombining aperiodic and voiced segments from initial stop-vowel syllables with different vowel contexts. Their findings were consistent with Cooper et al. (1952), Fischer-Jørgensen (1972) and many others which showed that <u>both</u> burst and formant transitions provide distinctive acoustic information for identifying place in a complementary or cue-trading relationship. Of particular interest to the present investigation was the finding that

formant transitions in these CV syllables appeared to be more important perceptual cues in some vowel contexts than in others. Dorman et al. also emphasized the importance of dynamic cues to place of articulation. That is, they observed that the burst was effective only when its main frequency peak lay close to the main formant of the following vowel such that the listener could track the changes in the articulatory gestures by following the changes in the spectral resonances. For example, the high frequency alveolar bursts were effective place cues before high front vowels, but not before low back vowels where formant discontinuities were present.

In contrast to the Dorman et al. study, Stevens and Blumstein (1978; Blumstein and Stevens, 1979; Blumstein and Stevens, 1980) have attempted to describe and experimentally verify the existence of static integrated acoustic cues for place of articulation in stops. They have argued that invariant acoustic properties for place can be found in the gross shape of the spectrum at the onset of the release burst. They claim that a unique spectral shape can be found for each particular place of articulation. Furthermore, they have argued that these spectral shapes are correlates of the phonologically distinctive features that define place of articulation and that they can be observed across syllable position, consonant manner class and talker (Jakobson, Fant and Halle, 1952).

Stevens and Blumstein have developed specific descriptions or templates of the gross spectral shapes for each place of articulation. They derived these templates from single 25.6 ms spectral sections taken at the onset of stop-vowel syllables and smoothed by linear prediction. Blumstein and Stevens (1979) then carried out several experimental tests to assess the power of these templates in identifying place of articulation for syllable initial and final stops produced in five vowel contexts by six talkers. Their results showed that the templates were fairly accurate for identifying the place of articulation of stops in syllable-initial position, but not of stops in syllable-final position.

Stevens and Blumstein then carried out two further studies to verify experimentally that the onset spectra were important perceptual cues for identifying place of articulation. Both studies used synthetic CV syllables (Stevens and Blumstein, 1978; Blumstein and Stevens, 1980). The synthetic stimuli varied in overall duration from relatively long CV syllables to very brief truncated stimuli, although they were all constructed using the same synthesis principles. These principles involved preserving the natural details of the burst, VOT and voiced formant transitions in their "full-cue" set. All these parameters were then manipulated to produce a stop consonant-vowel continuum from /b/ to /d/ to /g/ before three vowels. Subjects were able to identify some of these stimuli as

/b/, /d/ or /g/ on 100% of the trials in a forced choice task, while other stimuli were identified ambiguously. Stevens and Blumstein then observed, in an informal way, that the onset spectra for the unambiguously identified stimuli were in agreement with the proposed gross spectral shapes. From these findings they argued that the gross shape of the spectrum at onset contains the distinctive acoustic information to identify place of articulation in stops.

Although many other experimental conditions were included in their two reports, the basic strategy for verifying the role of gross spectral properties in synthetic CV syllables remained basically the same. By their own admission, these data "do not constitute a strong test of the theory" (Stevens and Blumstein, 1978, p.1367). Indeed, their experimental procedures appear to be unsatisfactory as a way of verifying whether the gross spectral shapes are sufficient perceptual cues for place of articulation, primarily because acoustic properties of the stimulus set were not manipulated in terms of the gross onset spectral properties themselves, but rather in terms of burst frequency, VOT and formant transitions. Nevertheless, Stevens and Blumstein argued that invariant acoustic cues for specifying place of articulation can be found in the first 10 to 20 ms of a stop waveform. These particular claims motivated several aspects of the present investigation.

Using quite a different approach to describe the acoustic cues to stop consonants, Searle and his colleagues have suggested that acoustic analysis techniques for speech signals should incorporate known psychophysical properties of the human auditory system (Searle et al., 1979; Searle et al., 1980). They have argued that the acoustic cues to linguistic contrasts should be described in terms of auditory representations that model the filtering effects of the peripheral auditory system. Searle et al. developed a frequency-by-amplitude analysis of speech signals based on analogue 1/3 octave filters, updated at 1.6 ms intervals. This approach thus emphasized not only the auditory transformation of the speech signal, but also the dynamic changes in the transformation. Their analyses permitted them to construct a three-dimensional representation, displaying the running spectra of a speech signal as it changed over time. In a preliminary study of the properties of these running spectra, Searle et al. were quite successful in identifying cues to voicing for stop consonants, but only partially successful in identifying cues to place. The idea of examining distinctive cues in running spectral displays provided the basis for developing a similar analysis technique in this investigation to study several long-standing questions concerning acoustic invariance in stop consonants.

Since the invariance issue plays an important role in motivating the present investigation, it will be helpful to

clarify the opposing positions on the topic adopted by, for example, Liberman and his colleagues, on the one hand, and by Stevens and Blumstein, on the other. For Liberman, the most basic property of human speech is that invariant cues cannot be found directly in the speech signal (Liberman et al., 1967). That is, speech is a dynamic, highly encoded representation of the phonemic structure of language -- a speech code. By contrast, Stevens and Blumstein (1980) claim that invariant acoustic cues can be specified for all the distinctive features of language. These two conflicting positions are revealed in differing approaches to several aspects of the invariance issue.

First, several sources of phonetic variation can affect the acoustic signal associated with a particular token of a stop consonant syllable. These factors include vowel context, syllable position, speaker differences (especially vocal tract size), rate of speech, and dialect. Liberman and his collaborators believe that an invariant acoustic cue would have to be some property or set of properties that remains invariant over all these sources of variation. In a recent paper, Liberman and Studdert-Kennedy (1978) have further elaborated the position that the encoding of acoustic cues throughout the syllable precludes the existence of physically invariant cues (cf. Liberman and Pisoni, 1977).

Although they argue that invariant cues can be specified, Stevens and Blumstein (1978; 1980) do not

clearly state what sources of phonetic variation are to be accounted for by their theory. In their template matching study, they examined three sources of variation: vowel context, speaker differences, and syllable position. Men and women were included as speakers, although little attention was directed to generalizing their templates to children's speech. Furthermore, the issue of speaking rate was never addressed even though the information integrated within a fixed 25.6 ms analysis window might be expected to change with differences in speaking rate. Thus, Stevens and Blumstein have failed to describe the relation between their proposed invariant acoustic cues and the major sources of phonetic variation known to influence the physical variation of speech signals.

A second issue concerns the nature of the possible invariant acoustic cues. In the published literature, acoustic cues have sometimes been defined as "simple" acoustic properties and other times as "relational" acoustic properties. A simple acoustic invariant is presumably absolute, immutable and immune to the sources of phonetic variation mentioned above. For example, Cole and Scott (1974a) have argued that aperiodic burst segments are simple invariant cues for place in initial stops. Liberman's position has been that most invariant cues, if they can be found in the stimulus, must be simple cues (Liberman et al., 1967; Liberman, 1970; Liberman and Studdert-Kennedy, 1978). On the other hand, Fant and

Stevens have proposed that relational acoustic properties may serve as invariant cues to stop consonants. While Fant does not specifically use the phrase "invariant cue," his theoretical position clearly states that relational properties are the important acoustic characteristics which directly specify phonetic contrasts (Fant, 1960; Fant, 1973). He has also suggested that these acoustic properties may even be relational across two phonetic segments. For example, the frequency shift in the compact spectrum of the velar stop burst is due to the influence of the following vowel context and is a predictable acoustic correlate of velar place of articulation.

Stevens and Blumstein (1978) have incorporated relational properties in their description of the gross shape of the spectrum at onset. These cues are relational because the spectral shapes refer to relative levels of high frequency energy versus low frequency energy, as opposed to absolute levels. For Fant and Stevens, these more complex relational properties can serve as invariant cues in speech perception because they are the predicted output of the reasonably stable underlying articulatory gestures (Fant, 1968; Stevens and Blumstein, 1980).

The investigation of the acoustic cues for place of articulation in this thesis was motivated by two contrasting positions concerning the use of the temporal dimension in descriptions of speech. The acoustic information for bursts, as well as Stevens and Blumstein's

onset spectra, has often been assumed to be present in single, static spectral sections having no time dimension. In contrast, the information for formant transitions has been thought to be contained in the time-variation of the first few spectral resonances preceding the steady-state vowel spectra. Moreover, a basic theme of Liberman and other investigators at Haskins Laboratories has been that the perceptually relevant linguistic information is encoded in the dynamic changes of the speech signal (Liberman et al., 1967; Liberman, 1970; Dorman et al., 1977; Bailey and Sumerfield, 1977; Fowler, 1980; Studdert-Kennedy, 1980).

The concept of static spectra serving as acoustic cues in speech perception seems, in some respects, counter-intuitive. Rapid variation of the articulatory gestures and the associated acoustic signals generated in the vocal tract is one of the most distinguishing characteristics of human speech (see Liberman, 1970; Studdert-Kennedy, 1977; and Stevens, 1980). Furthermore, it is known that neural signals produced in the peripheral auditory system vary synchronously with variation in the input speech waveform (Kiang, Eddington and Delgutte, 1979; Kiang, 1980; Delgutte, 1980). Therefore, the proposal that static spectral sections can serve as perceptual cues for speech perception is at odds within the ecological framework of the continuous, rapid time variation inherent in the articulatory, acoustic and neural domain of speech signals.

Another factor influencing the course of this investigation was a technological one. We have developed several interactive digital signal processsing techniques for examining the fine details of the spectral and temporal properties of speech signals. Based on the linear prediction algorithms of Markel and Gray (1976), the computer program SPECTRUM was designed and implemented to calculate and display various spectral properties of speech (see Appendix A). The development of these analysis techniques was an important aspect of the overall research program.

Four experiments will be described. They were designed to examine the acoustic cues for place of articulation in initial stop consonants. A natural speech data base was also developed for this research which contained enough stop-vowel syllables from several talkers to examine thoroughly the acoustic cues in question.

The purpose of Experiment 1 was to reexamine the information contained in the formant transitions of natural, voiced stop consonants for possible cues to place of articulation. Although formant transitions were defined as changes in spectral resonances over time, previous studies had shown that formant transitions were sufficient cues to place of articulation for only a few vowel contexts in natural speech (Dorman et al., 1977; Ohde and Scharf, 1977). Furthermore, a detailed measurement study of formant transitions using modern digital techniques was not

available in the literature. The rationale behind Experiment 1 was to use the power of linear prediction analysis in the SPECTRUM program to measure carefully the parameters of formant transitions as possible correlates of place of articulation. In this study, we hoped to determine if spectral change over time, as measured in formant transitions, could provide reliable information for distinguishing place of articulation in initial voiced stop consonants.

In Experiment 2, we examined the acoustic correlates of place based on the other premise concerning burst and formant transitions, namely, that these properties should be treated as a unitary acoustic event. In contrast to the views of Stevens and Blumstein, we focused our attention on changes in the spectral distribution of energy over time in initial waveform segments. This goal was accomplished by implementing running spectral displays as suggested by Searle et al. (1979) in the SPECTRUM program. The linear prediction running spectra permitted us to display continuous changes in spectral energy from the release burst into the formant transitions. The purpose of Experiment 2 was to examine the running spectral displays of a large number of stop-vowel syllables and to develop a set of possible visual features which could be used to specify the correlates of place of articulation. A further goal of this experiment was to determine how some of the known psychophysical properties of the human auditory

system could be incorporated in these running spectral displays as Searle et al. (1980) have suggested.

The last two experiments were designed to verify several of the observations made in Experiment 2 and to compare these with the earlier claims made by Stevens and Blumstein (1978; Blumstein and Stevens, 1979; Blumstein and Stevens, 1980). Experiment 3 was carried out to test the hypothesis that sufficient acoustic information for identifying place of articulation is located in the initial portions of a stop-vowel waveform. Blumstein and Stevens (1980) claimed that only 10 to 20 ms of a stop-vowel waveform was necessary to identify place correctly from a small set of synthetic CV syllables. In our experiment, subjects identified the consonant from a larger number of natural truncated CV syllables produced by two talkers. The purpose of this experiment was to determine the duration of the initial waveform portion that is necessary to achieve relatively high levels of consonant identification. In a second part of this experiment, subjects also identified the vowel in these truncated syllables. The purpose of collecting both consonant and vowel identification responses was to determine if consonant and vowel information is encoded throughout the syllable as Liberman has and his colleagues have argued (e.g. Liberman et al., 1967; Liberman, 1970).

The final experiment sought to establish whether our time-varying spectra or the static onset spectra of Stevens

and Blumstein are better representations of the acoustic cues for place of articulation in initial stops. Two sets of truncated synthetic CV stimuli were designed to incorporate either the time-varying or the static onset spectral properties of stop-vowel syllables. Subjects identified the consonant in these brief stimuli and indicated how confident they were of their responses. By using both types of response, we hoped to determine what properties of these stimuli were used by subjects in identifying place of articulation. In particular, the outcome of this study should help to determine if the acoustic information contained in either the time-varying stimuli or the static onset spectra stimuli serve as the major perceptual cues to distinguish place of articulation.

In summary, the unifying theme of all four experiments is that change in the spectral distribution of energy over time is a proper characterization of the way the peripheral auditory system transduces speech signals and, therefore, that spectral change should be an inherent part of the descriptions of acoustic cues for place of articulation. The significance of this research is that progress in the long search for the invariant cues for place of articulation may be advanced by incorporating known properties of the auditory system, particularly the concept of change in the spectral distribution of energy over time in the analysis of the initial portions of stop-vowel waveforms. These findings should have important

implications for current theories of speech perception in terms of modeling the neural representation of speech signals at the earliest stages of perceptual analysis.

II. EXPERIMENT 1. FORMANT TRANSITIONS OF NATURAL STOP CONSONANTS

A. Introduction

The main purpose of Experiment 1 was to reexamine the acoustic information in formant transitions for possible cues to place of articulation in stop consonants. Previous research on the topic has been limited by the poor resolution of rapidly changing spectral information in sound spectrograms, particularly at the onsets of formant transitions. Formant transitions were measured in this study by digital signal processing techniques. By choosing analysis methods which would permit fine temporal resolution, more detailed information concerning the formant transition patterns could be obtained than was available from previous studies. Furthermore, with digital techniques, the role of changing spectral information as a possible cue to place is not limited to voiced formant transitions. The speech tokens collected for Experiment 1 are reexamined in Experiment 2 to observe changes throughout the voiceless spectra of release burst and aspiration continuously into the voiced spectra.

Although considerable attention has been directed to formant transitions as cues to place of articulation in stop consonants, especially by researchers at Haskins Laboratories (e.g., Liberman et al., 1967), very few studies have actually measured parameters of formant

transitions in natural speech. The earliest research on the sufficiency of transition cue information came from studies of synthesized speech at Haskins. These studies examined the role of F2 transitions (Liberman et al., 1954), F3 transitions (Harris, Hoffman, Liberman, Delattre and Cooper, 1958), and the derived concept of formant locus (Delattre, Liberman and Cooper, 1955) in stop consonants synthesized on the Pattern Playback.

Partly in response to results from the synthetic speech research, two studies of natural speech formant transitions were conducted. The first was a general study of formant movement throughout CVC syllables reported by Lehiste and Peterson (1961). This report emphasized the lack of success in obtaining constant F2 onset frequencies and formant durations for each place of articulation. However, detailed formant measurements from their very large corpus of CVC syllables were not presented. In the second study, Öhman (1966) examined the coarticulation effects on formant transitions for stops in VCV utterances, primarily in Swedish. Like the earlier Lehiste and Peterson (1961) study, this research was not a parametric investigation of formant transitions. However, in Öhman's report formant loci for F2 and F3 were calculated according to rules spelled out in the earlier study by Delattre et al. (1955). Öhman showed that invariant formant loci for F2 and F3 were not obtained from measurements of transitions in natural speech.

One very detailed study of the spectral and durational acoustic parameters for stops in CV syllables was reported by Fant (1973). His data consisted of spectrograms of six initial stops combined with the nine long vowels of Swedish spoken once by one talker. Fant's report contains a great deal of information, not only about formant frequency onsets and steady-states, voice onset time, and durations of transitions, but also on how to apply this information to synthesis of CV syllables. Based on these results, Fant (1973) offered two major conclusions. The first, in agreement with Öhman's (1966) study, was that a constant locus for place could not be obtained from the spectrographic measurements. Second, F2 and F3 formant transition patterns did not uniquely specify place of articulation in his or Öhman's data. The dynamic formant patterns for many CV pairs, e.g. /ko/ - /po/, were about the same. Thus, Fant concluded that formant transition patterns alone are not sufficient correlates of place of articulation in stops. Fant then reemphasized his earlier hypothesis (from Fant, 1960) that place cue information is located in the combination or integration of burst and formant transition information in the first 10 to 30 ms of a stop waveform. This hypothesis, which is in close agreement with the recent views of Stevens and Blumstein (1978), will be examined in greater detail in Experiments 2, 3 and 4.

Given the relatively sparse masurements of natural formant transitions in the literature, we decided to carry out a detailed parametric study of formant transitions in English stop-vowel syllables using modern digital signal processing techniques. The goal of this experiment was to examine enough tokens from a single talker to determine whether any of the parameters measured could provide acoustic information to reliably specify correlates of place of articulation. It was the intent of this study to examine an exhaustive list of formant transition parameters. Although research over the years provided little support for the presence of invariant cues to place in the formant transitions, the present study attempted to determine which of several parameters of formant transitions could provide information for specifying place of articulation in stop consonants.

B. Data Base

The syllables analyzed consisted of the voiced stops /b, d, g/ followed by /i, I, e, ɛ , ae, a, o, u/. Only voiced stops were included in the set because the purpose of this study was to measure voiced formant transitions as traditionally defined, and the aspirated stops in English have minimal voiced transitions.

The 24 test syllables were embedded in the carrier sentence, "Teddy said CV." The sentences were presented under computer control on a CRT display monitor in a sound

attenuated room (IAC Model 401A). One male speaker of American English, a phonetician, recorded 10 randomizations of the sentences on an Ampex AG-500 taperecorder using an Electro-Voice D054 microphone. Five of these lists, beginning with the third list, were digitized and stored on disk. The waveforms were first low-pass filtered at 4.9 kHz and then sampled at a 10.0 kHz sample rate using a 12 bit A/D converter with a PDP-11/05 computer. The sentences were edited so that only the target CV was permanently stored on disk.

C. Analysis Procedure

Formants were calculated using the SPECTRUM program (see Appendix A) written by the author and based on the linear prediction algorithms of Markel and Gray (1976). The linear prediction coefficients were calculated using the autocorrelation method. The formant transitions for all syllables were determined using the interactive graphics of SPECTRUM in conjunction with strict guidelines for determining the onset of voicing and the vowel steady-state as described below.

The first 95 ms of each CV was analyzed every 5 ms. Each waveform segment, or frame, was preprocessed by both first differencing (pre-emphasis filtering) and application of a 200 point (20 ms) Hamming window. The shape of a Hamming window is similar to that of one half-period of a sine wave. Multiplication of the speech signal by a Hamming window reduces spectral distortion in the calculated output

spectrum (see Markel and Gray, 1976, p. 157). The Hamming window for the first frame was carefully positioned by hand editing so that the burst onset was at the center of the window. Fourteen linear prediction coefficients were then calculated for each frame except for a few syllables where formant tracking was substantially improved by using 12 coefficients.

Formant trajectories were calculated for the first four formants (F1 to F4) for the first 95 ms of each CV syllable. For this speaker, F4 was not consistently detected for some CV's, and these syllables were not reanalyzed to enhance F4. It was decided, however, that the data base for this study should include only CV's where F1, F2 and F3 were consistently tracked by the current algorithms. Therefore, when occasionally one formant was not detected, analysis parameters were changed, for example, by changing the number of analysis coefficients from 14 to 12 or 16. If three formants still could not be detected, then another test sentence from one of the previously undigitized lists was substituted. A total of four sentences were substituted in the entire analysis.

Guidelines were also developed and strictly adhered to for measuring formant transitions. A formant transition was defined as the segment of a formant beginning at the onset of voicing and ending at the onset of the steady-state portion of the vowel. The first step was to determine the frame in which voicing began. Spectral sections were

examined for each 5 ms frame at the beginning of each CV in conjunction with the value of RMS energy for each frame. The two-fold criteria for extablishing voicing onset was the presence of a sharp F1 spectral peak with evidence of peaks for F2 or F3, accompanied by an abrupt increase in RMS energy. The values for the onset frequencies and onset frame for F1 to F4 were recorded from the printout of the trajectories. F1 always onset in this frame, but the higher formants sometimes started in later frames.

Insert Figure 1 about here

Vowel steady-states were determined from CRT displays aligned in time with the waveform (see Fig. 1). There are many possible definitions for vowel steady-state. In this study, criteria were adopted for determining the onset of the steady-state portion of the vowel for each formant individually. The onset of the steady-state began in the frame where the frequency change fell to less than 10 Hz per 5 ms frame, or where the formant back-tracked to the same value within 4 frames. The measurement procedure was facilitated by visual feedback from the CRT display using a manually controlled cursor to mark the frames. Figure 1 also shows typical markings for the steady-states in the syllable /da/. In some cases, particularly with /g/, another five frames (120 ms total) were analyzed in order to determine the steady-states. In a few other cases,

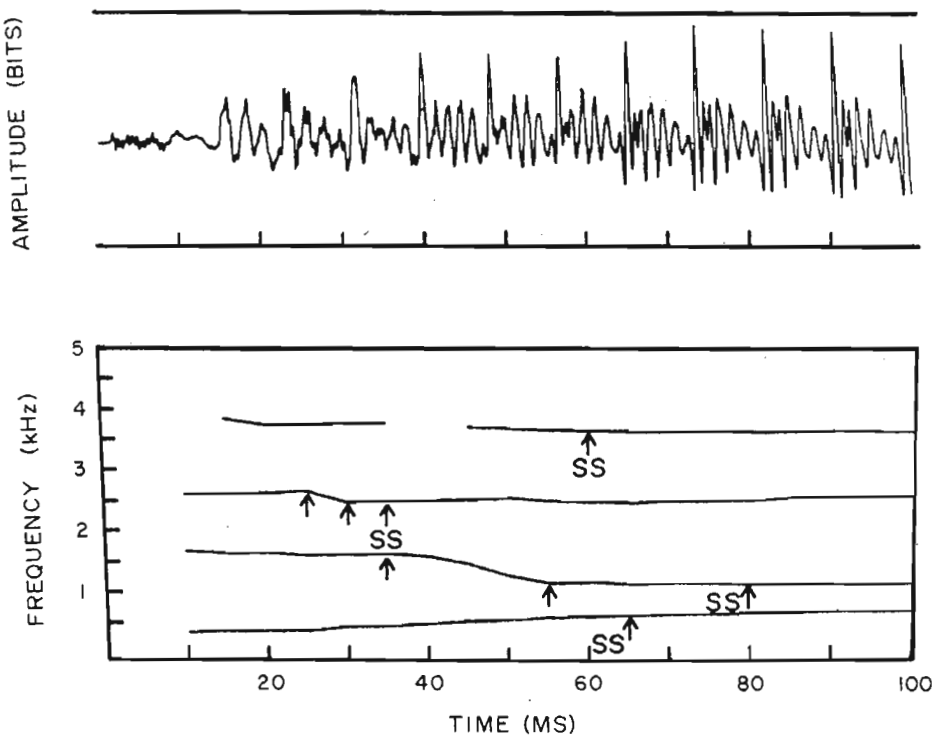


Figure 1. The SPECTRUM CRT display of the waveform aligned with the formant trajectories was used to determine the formant transitions. In addition to the onset values of the transitions, the values measured for each of the four formants of /da/ are shown by arrows. "SS" notes where the vowel steady-state began. "↑" notes where there were deviations from a linear approximation of the formants transition.

formant trajectories were flat and therefore had no discrete transition segment. The frequency and frames for the onset of the steady-state were recorded for all four formants for each token.

Although formant transitions are often approximated by straight line segments in synthesis experiments, we also sought to determine if straight line approximations were in fact reasonable characterizations of formant transitions in CV syllables. After the steady-state frames were located, the formant transitions on the CRT displays were examined to determine if distinct deviations from linearity could be observed. All the segments on which a peak, dip or flat portion of the transition was observed were noted in the data tables. Data tables also contained the frequency values for each formant 95 ms from the burst. Table 1 provides an example of a formant transition table.

Insert Table 1 about here

From these tables, each CV was plotted on graph paper, three CV's per graph, grouped as /b,d,g/ for a given vowel for one list randomization. Figure 2 is a display showing the formant transitions for /bI, dI, gI/ recorded in list P21. The data tables and the 40 graphs of the individual tokens comprised the basic data base for Experiment 1.

Table 1. Formant transition data for /de/ from list P21. "SS" means steady-state.

	Onset Frame	Onset Hz	Frame	Hz	Frame	Hz	95 ms from burst (Hz)
F1	3	334	12	467(SS)			487
F2	3	1814	flat				1883
F3	3	2591	5	2635(peak)	8	2551(SS)	2547
F4	3	3849	9	3604(dip)	10	3718(SS)	3378

Insert Figure 2 about here

D. Results

The first step in the analysis was to establish if each formant transition could be approximated by a single straight line segment. The formant transitions for each of the five repetitions of a CV were compared visually. The data collected for F4 were deleted from further analysis because approximately 26 percent of the F4's could not be tracked by the SPECTRUM program, and the remaining F4's showed considerable token to token variability.

Comparisons of the first three formant transitions revealed that they could all be approximated by straight lines except for F3 in the syllable /do/. F3 for /do/ was approximated by three straight lines because there was a dip in the transition approximately 20 ms after voicing onset (see Fig. 3d). This dip can be predicted from the change in cavity affiliation for F2 and F3 for /d/ before rounded back vowels (Fant, 1960, sec. 1.4). For this speaker, however, this change was noticable only in the voiced portion of the transitions for F3 in /do/. It was therefore decided that the transitions for F1, F2, and F3 could be modeled as straight lines from their onset frequency values to the steady-state frequencies of the vowels, and that F3 for /do/ would be treated as a special case in the analysis where necessary.

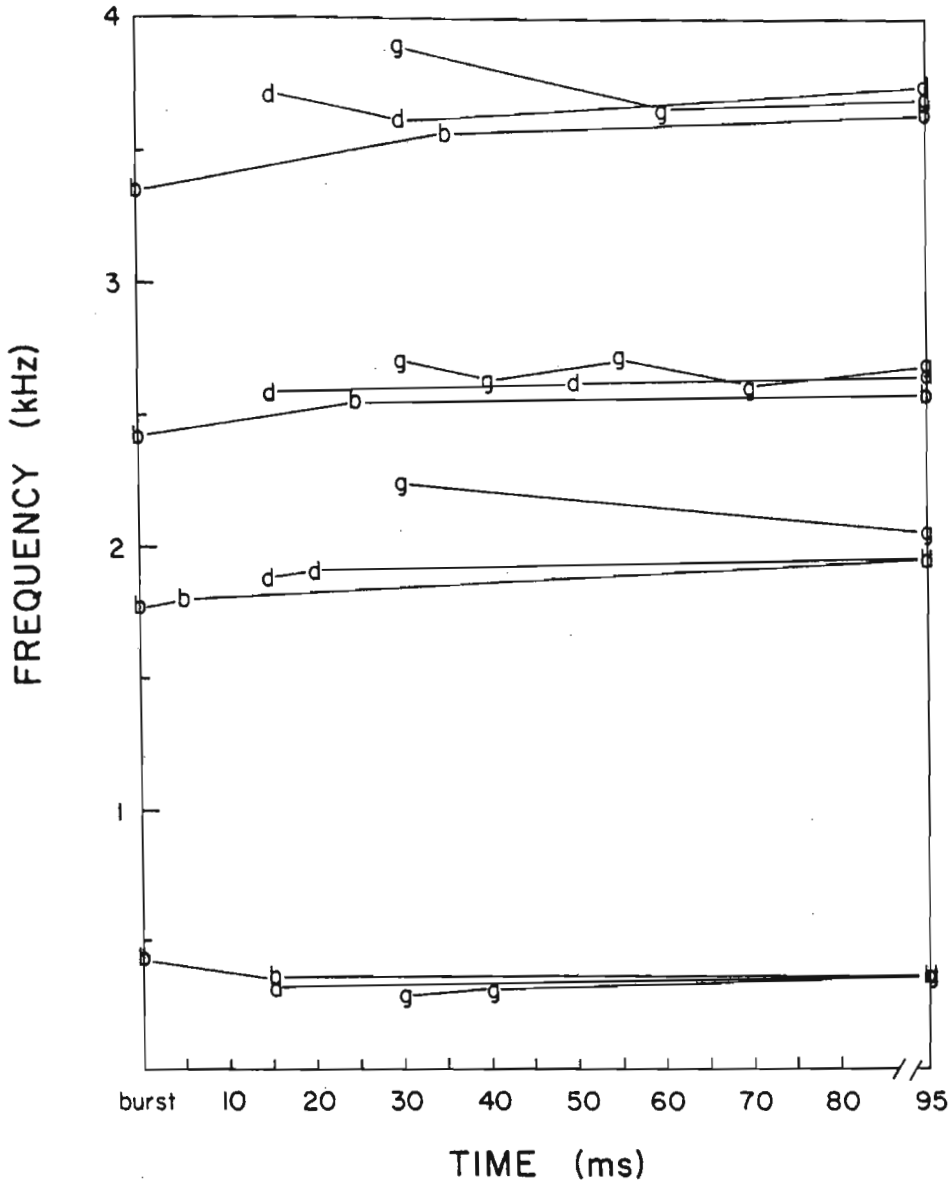
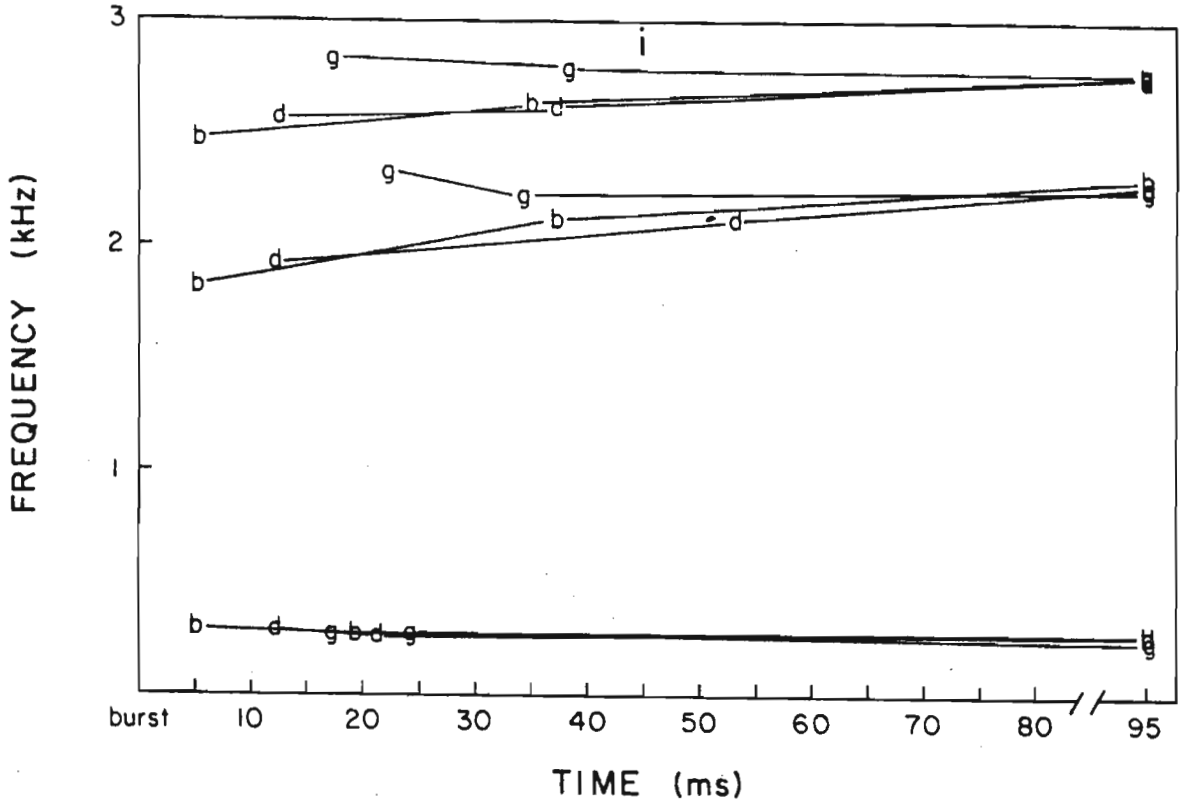


Figure 2. Sample data graph of the formant trajectories measured for individual tokens in Experiment 1. Shown here are /b,d,g/ with /I/ on list P21.

In order to determine whether formant transitions can reliably encode distinctive information for place of articulation, an extensive series of statistical analyses was performed on these acoustic measurements. The measurements of the onset and steady-state frequencies and durations of the linear transition segments for F1, F2 and F3 were averaged and subsequently analyzed by a one-way analysis of variance with Scheffé post hoc analyses. All statistical analyses reported here were calculated using the Statistical Package for the Social Sciences Programs (Nie et al., 1975). Average values for the parameters of the formant transitions measured in this study are displayed in Figures 3a - 3d and the numerical values are given in the data tables in Appendix B.

Insert Figure 3 about here

Liberman et al. (1954) claimed that the properties of formant transitions that serve as perceptual cues to distinguish place of articulation are the duration and extent of the transition segments. This description typifies the Haskins point of view that dynamic changes in frequency specify the perceptual cues to stop consonants. In a further report on formant transitions, Delattre et al. (1955) noted that if the transitions for /b,d,g/ before the vowel converge to the same vowel formant frequency value, the description of the transitions can be simplified. In



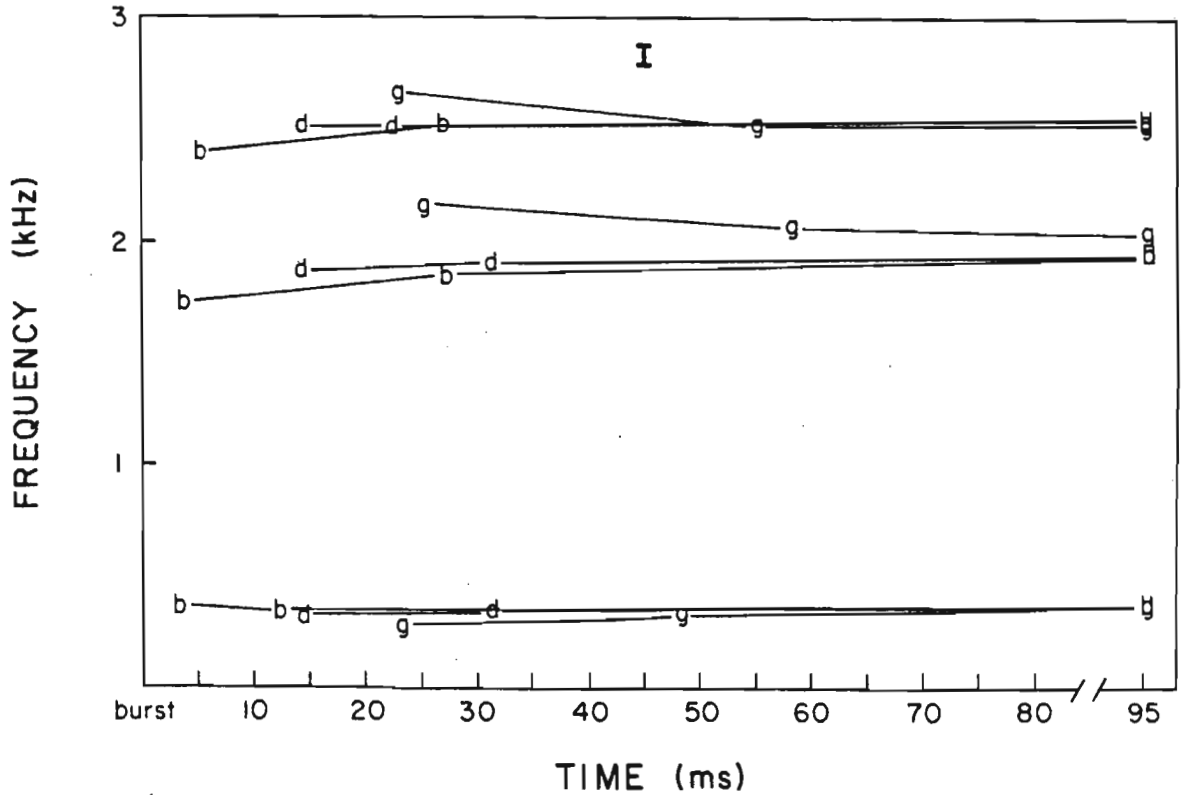
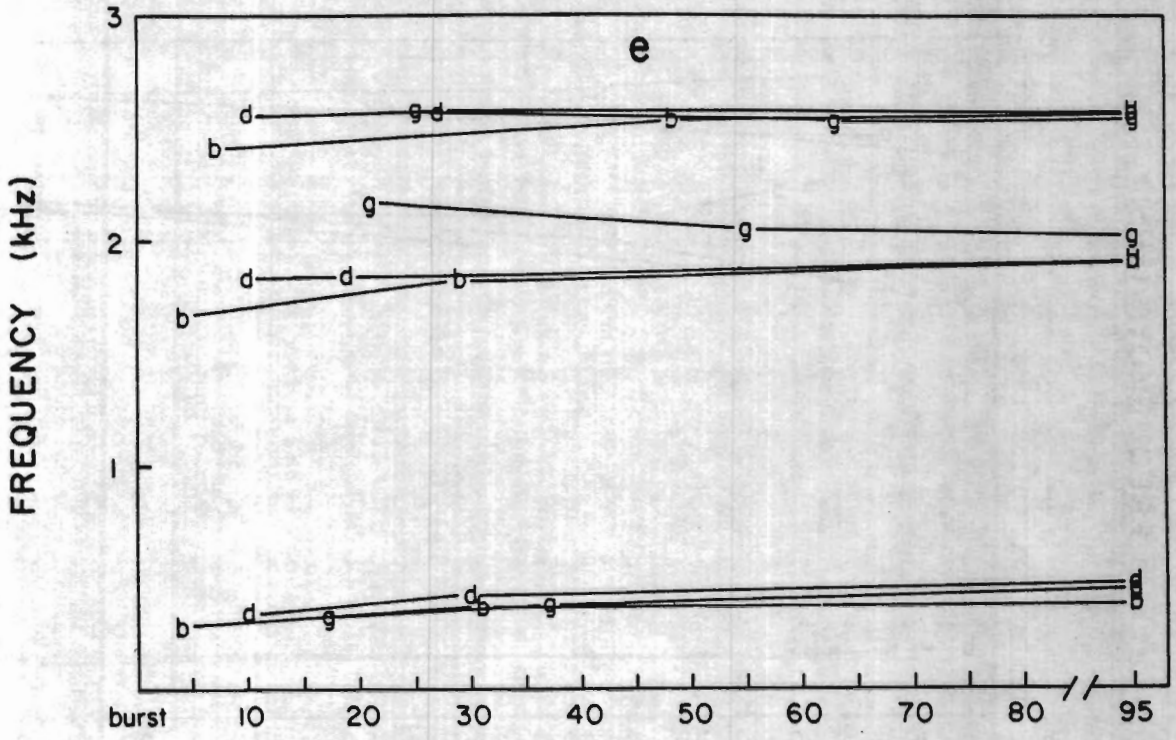


Figure 3a. Average values of the formant transitions for /b,d,g/ with /i/ and /I/.



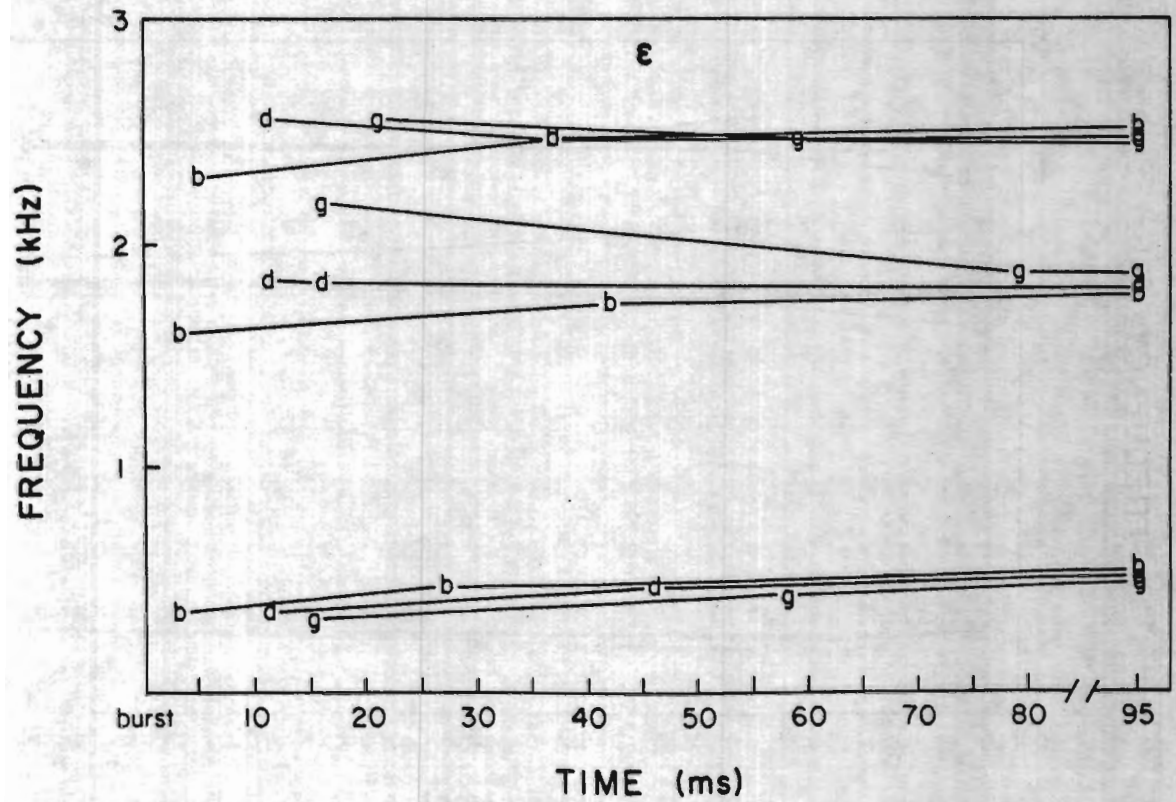
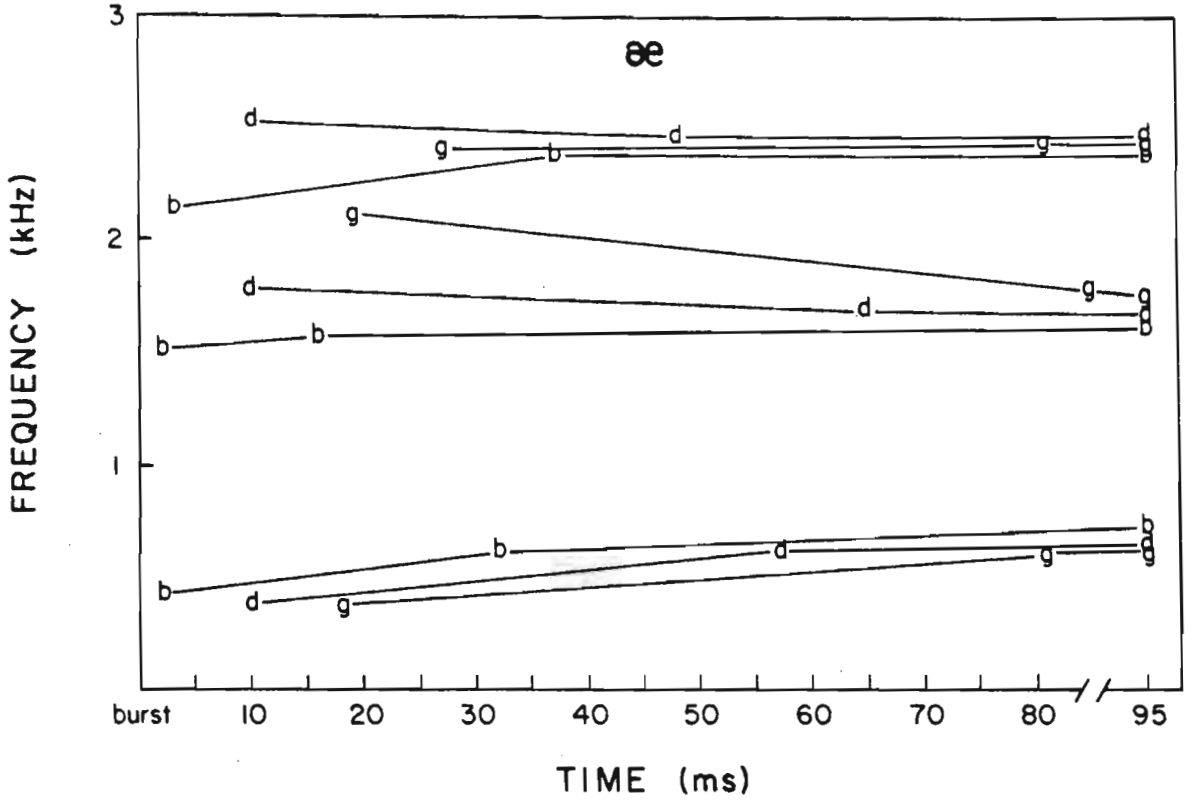


Figure 3b. Average values of the formant transitions for /b,d,g/ with /e/ and /ε/.



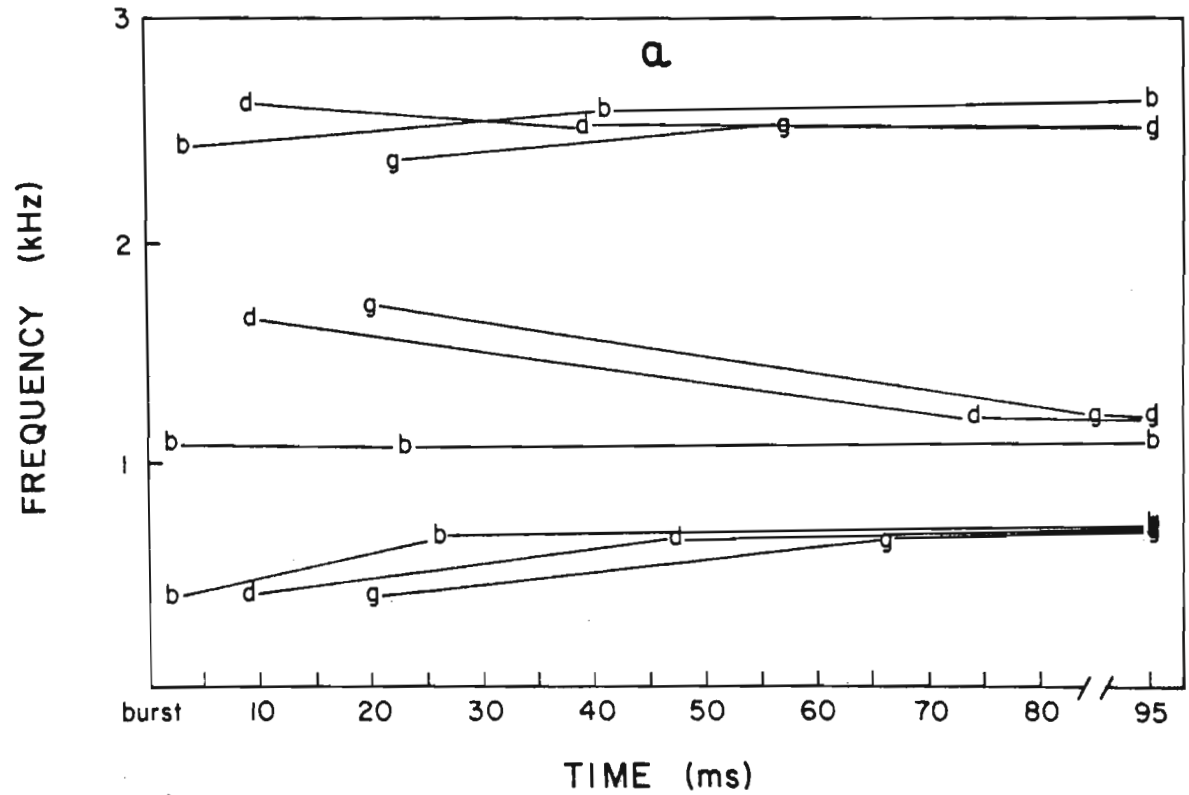


Figure 3c. Average values of the formant transitions for /b,d,g/ with /ae/ and /a/.

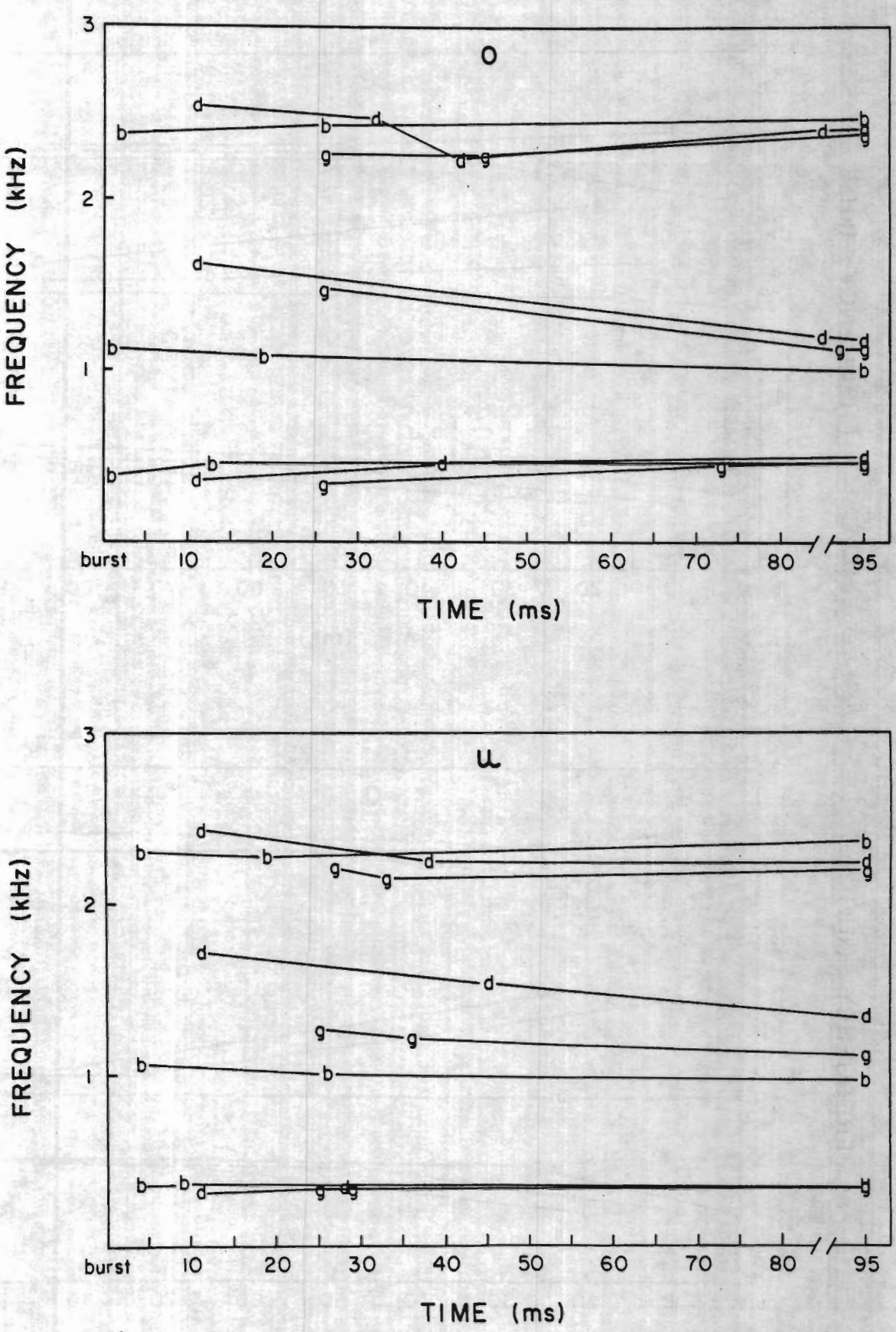


Figure 3d. Average values of the formant transitions for /b,d,g/ with /o/ and /u/.

particular, a formant transition can be described by measuring its onset frequency, its steady-state frequency, and the duration of the transition segment. Following this suggestion, we measured the steady-state frequencies and analyzed them to see if they were distinctively different for the preceding consonant context. As we shall see below, the values of the steady-state vowel formants were not statistically different, so most of the remaining analyses concentrated on onset frequency and durational properties of the formant transitions.

The first question examined in this study was whether the steady-state values of the vowel formants were the same for all consonants. Table 2 summarizes the results of the consonant groups produced by the Scheffé post hoc analysis.

Insert Table 2 about here

Although the one-way analysis of variance was significant in 12 out of 24 analyses, distinctive groupings of the formants were not obtained for consonantal context except in one case. Only the F2 for /u/ showed significant differences in the steady-state portion of the vowel between /b/, /d/ and /g/. Thus, the vowel targets produced by this speaker were quite similar across the initial stops, demonstrating that vowel steady-state frequencies alone could not serve as reliable correlates of place for /b,d,g/.

Table 2. Statistical groupings from the Scheffé post-hoc analysis of steady-state formant frequencies for each vowel formant. Stops inside parentheses indicate consonantal contexts which were not statistically different. ### indicates analysis separated /b/, /d/, and /g/ into distinct groups.

	F1	F2	F3
i	(bd)(dg)	(bdg)	(bd)g[a]
I	(bdg)	(bd)g[a]	(bdg)
e	(bd)(dg)[a]	(bd)g[a]	(bdg)
ɛ	(bdg)	(bd)(dg)[a]	(bdg)
ae	(bdg)	b(dg)[a]	(bdg)
a	(bdg)	b(dg)[a]	(bdg)
o	(bdg)	(bg)d[a]	(bd)g[a]
u	b(dg)[a]	###[a]	(bd)g[a]

[a] $p < 0.01$

Next we considered if the onset frequencies of the formant transitions were distinctive for different places of articulation. Generally it has been assumed in the literature that the transition onsets serve as highly distinctive acoustic cues for place of articulation, although they are conditioned by the following vowel context. However, we decided to test the possibility that transition onsets were invariant over vowel contexts by averaging F1, F2, F3 onset frequencies across all eight vowels. A one-way analysis of variance revealed significant differences for all three formants at better than .001 level. The Scheffé analysis revealed, however, that stops were always grouped as /b/ alone and /d,g/ together for all three formants. In particular, there was a tendency for the F1 onset frequencies of /b/ to be higher than those of /d/ and /g/, and the onsets of F2 and F3 of /b/ to be lower than those of /d/ and /g/. As we shall see below, however, this is only a statistical tendency and the findings do not successfully classify place of articulation into three distinctive categories.

Insert Table 3 about here

How distinctive were the formant transition onsets for /b,d,g/ when the vowel context was known? Figure 3(a-d) shows the average formant frequencies plotted by vowel context. Table 3 summarizes the results of the statistical

Table 3. Statistical groupings from the Scheffé post-hoc analysis of onset formant frequencies for each vowel formant. Stops inside parentheses indicate consonantal contexts which were not statistically different. ### indicates analysis separated /b/, /d/, and /g/ into distinct groups.

	F1	F2	F3
i	(bdg)	(bd)g[a]	(bd)g[a]
I	b(dg)[a]	###[a]	b(dg)[a]
e	b(dg)[a]	###[a]	(bdg)
ɛ	(bdg)	###[a]	b(dg)
ae	b(dg)[a]	###[a]	###[a]
a	(bdg)	b(dg)[a]	(bg)d[a]
o	(bdg)	###[a]	(bg)(bd)[a]
u	(bdg)	###[a]	(bdg)

[a] $p < 0.01$

analyses. The results show that F1 alone cannot be used as a cue for place since it was found to separate /b/ from /d/ and /g/ for only three vowels in the set and it showed no significant differences for place for the other vowel contexts. F2 onsets were significantly different for place for all eight vowels. However, the Scheffé analysis revealed differences in the place categories for only six of the eight vowels (all but /i/ and /a/). For F3, onsets were significantly different for five vowels, but only one vowel /ae/ showed distinctive categories for /b/, /d/ and /g/. Thus, complete separation of /b,d,g/ onsets for both F2 and F3 was obtained for only one vowel, /ae/. On the other hand, for two vowels, /i/ and /a/, neither F2 nor F3 onsets distinctly separated place. For the remaining five vowels, F2 onsets distinguished place. However, this was not observed for differences in F3. Therefore, the statistical analyses performed here did not successfully categorize place of articulation for /b,d,g/ for all vowel contexts. These results suggest that, even on a statistical basis, F1, F2 and F3 onset frequencies each taken separately cannot be used to differentiate place of articulation in initial stops.

Durations of the formant transitions were also calculated, averaged and analyzed as possible correlates of place separately for each vowel. Transition durations varied substantially from token to token such that the standard deviations were extremely large. The statistical

analysis showed that only the F1 transition duration for the vowels /ae/ and /o/ could successfully sort the stops for place. Thus, for the vowels produced by this talker no consistent differences in transition duration distinguished among places of articulation.

These results can also be interpreted in a manner closer to the Haskins point of view. Since the steady-state values for a vowel formant were undifferentiated, the direction and extent of a formant transition can be determined from the onset frequency and transition duration measurements. However, the transition durations were also not reliably differentiated across consonant contexts. Thus, for a given vowel, the direction and extent of a formant transition is derived from its particular onset frequency, and values of transition duration and vowel steady-state averaged across consonant context. But our results have already shown that the F1, F2 and F3 onset frequencies alone did not differentiate place of articulation. Thus, taken togther, our results imply that the direction and extent of the formant transitions measured from natural CV syllables cannot be used as reliable cues to distinguish place of articulation for <u>all</u> vowel contexts. This point is illustrated clearly in Fig. 3 where formant transitions for the syllables /bi/ and /di/ are nearly the same.

Insert Figure 4 about here

Perhaps, a combination of F2 and F3 information would be more useful. One way to represent the combined place information contributed by F2 and F3 onset frequencies is to display them in an F2 X F3 space. Figure 4 shows the formant onsets plotted in a linear F2 X F3 space. Discriminant analysis was used to calculate an optimal F2 X F3 space separately for each vowel. Stops were correctly classified for place for 97% of the syllables. One /bi/ was misclassified as /di/, one /bI/ as /dI/ and one /do/ as /go/. Thus, the combined information contained in the onsets of both F2 and F3 taken together are effective in distinguishing place of articulation when the vowel context is known. To see whether the F2 X F3 onset space provided context invariant cues for place, a Discriminant Analysis was performed across all eight vowels. Place was correctly classified for only 68% of the syllables in the corpus. Thus, the combined information of the onsets of F2 with F3 considered over all vowel environments does not appear to provide context invariant cues for place.

Another way to characterize invariant information for place of articulation was suggested by Delattre et al. (1955) in terms of the concept of formant locus. The rationale underlying the definition of formant locus was that since the articulatory closure for each of the labial,

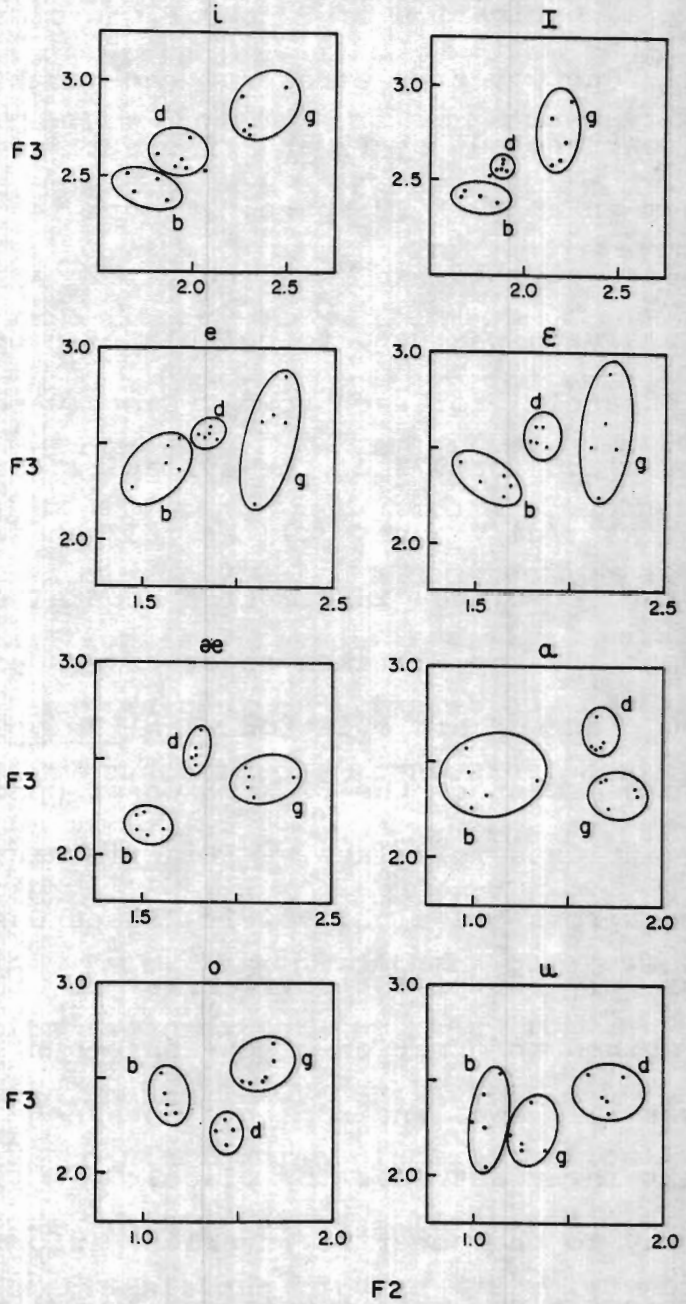


Figure 4. Each panel shows the groupings of the formant frequency onsets produced by Discriminant Analysis for a given vowel. Frequency is in kHz.

alveolar and velar consonants was relatively fixed across vowels, there should be an invariant formant starting frequency or locus for each place. This frequency would not be radiated acoustically due to the inevitable shift in formant frequencies with release of articulatory closure, but might be extrapolated backwards from the actual formant transitions. Thus, for a given consonant, the formant transitions moving from fixed loci to the appropriate vowel steady-state frequencies represent a direct acoustic analogue of the underlying articulatory gestures (see also Stevens & House, 1956).

The loci for F2 and F3 for each CV were calculated as follows. Examining Fig. 3 it can be seen that the time of onset of the voiced formant transitions relative to the burst is quite variable across consonants, and even across formants within the same CV syllable. We adopted the suggestions of Stevens and House (1956) that the acoustic locus should be referenced to the point of articulatory closure. In this analysis, articulatory closure was represented as the time of burst release which was the common point of alignment used in the digital analysis. The locus for any formant was therefore determined by extending the voiced formant transition backward in time to the burst. In this case, the projected formants were extended as linear segments with the same slope as the voiced formant transitions.

Specifically, the locus for each F2 and F3 was calculated in the following way. The slope of the formant was calculated for the formant transition from its onset to the steady-state portion of the vowel. Since the time of onset of the transition relative to the burst had been recorded earlier, the standard equation for defining a straight line, y = slope*x + intercept, was used to calculate the locus frequency at the burst, which represented the intercept. The slope, locus and projected frequency 10 ms from the burst were calculated separately for F2 and F3 for all 120 syllables, and then averaged for each of the five repetitions of each syllable type. These values are plotted on Figure 5 by consonant and vowel. The circles indicate the loci of the formant transitions. The line segment radiating from a locus represents the first 10 ms of the projected formant transition.

Insert Figure 5 about here

For the labial place of articulation, the loci for F2 and F3 clustered into two groups, one for front vowels /i,I,e,ɛ,ae/, and one for the back vowels /a,o,u/. For the front vowels, the loci shifted downward for both F2 and F3 as the formants specifying the following vowel decreased in frequency. The loci for the back vowels were more tightly clustered. These results for labial stops indicate that, at best, there are two F2 loci for /b/, one at about 1645 Hz

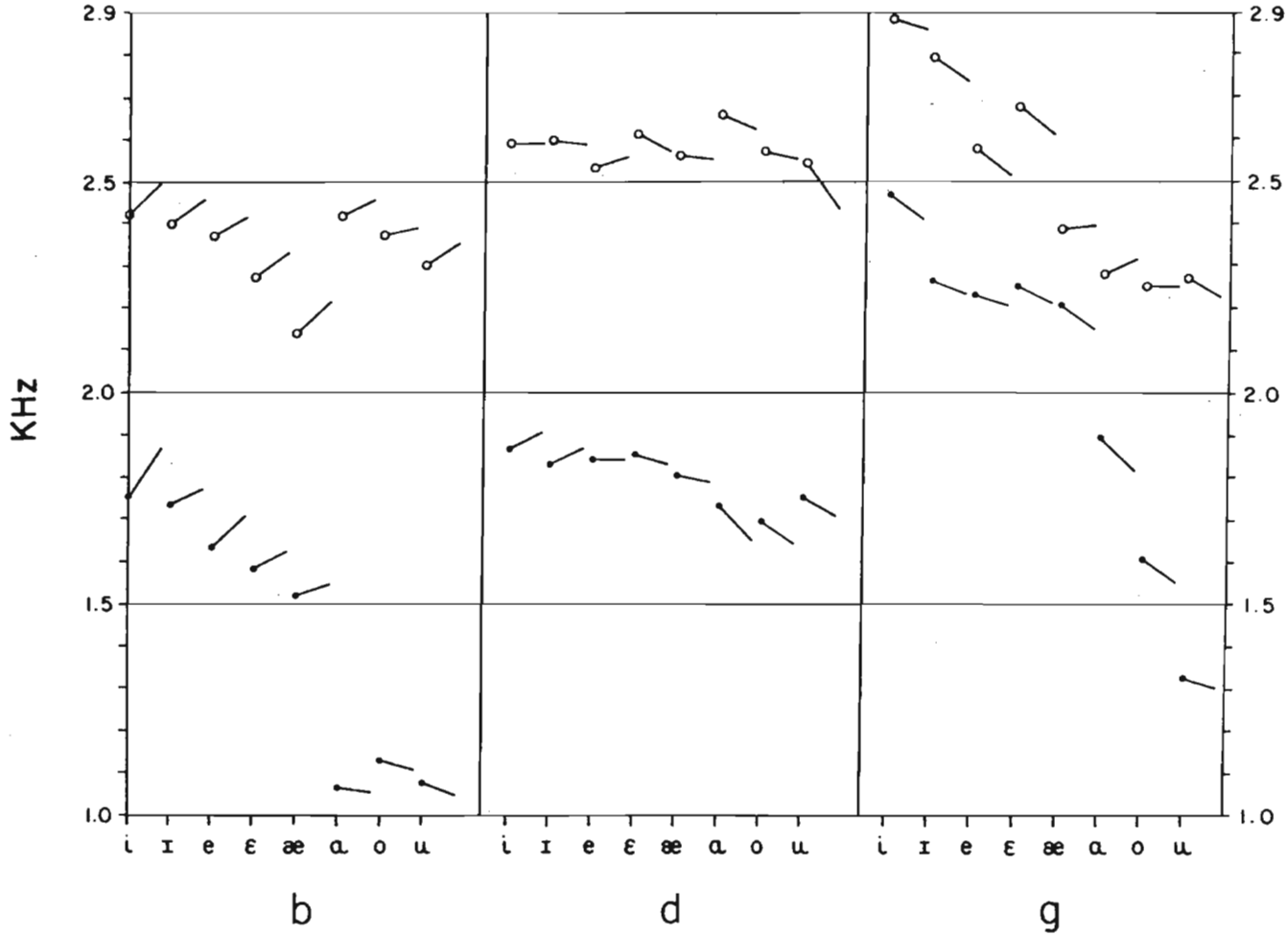


Figure 5. Formant transition loci for F2 (solid circles) and F3 (open circles). Line segments represent first 10 ms of projected formant transitions.

averaged over front vowels and one at about 1090 Hz averaged over back vowels. For F3, one locus at about 2337 Hz can be derived from the average across all vowels.

For the alveolar place of articulation almost invariant loci emerged for both F2 and F3. The average value for F2 was 1797 Hz, which is very close to the 1800 Hz value reported in the original Delattre et al. (1955) synthesis study, as well as in the Stevens and House (1956) acoustic modeling study. For F3 the average value was 2581 Hz.

The loci for the velar stops showed little evidence of clustering for F2 or F3. Instead, F2 and F3 loci showed a downward frequency shift from /i/ to /u/. This shift corresponds to the well-known change in place of articulation of English /g/ from palatal before /i/ to velar before back vowels.

Associated with the concept of a fixed acoustic locus for place of articulation is the notion that the formant transition slopes have relatively fixed patterns, particularly for /b/ and /g/. The slopes can be seen in Fig. 5 as the first 10 ms of the projected formant transitions. In the literature, F2 and F3 for labials are generally said to rise (Cooper, et al. 1952; Stevens, 1975; Stevens and Blumstein, 1978). For this speaker, however, F2 falls for back vowels. For alveolars, transition slopes are not considered to have an invariant pattern, a finding that was consistent with the present data.

The concept of "compact" has been used to describe the distribution of frequencies for velar consonants. While Jakobson et al. (1952) and Fant (1960) have referred to a compact burst spectrum, Stevens (Stevens, 1972; Stevens and Blumstein, 1978) has interpreted compact to mean that the F2 and F3 loci are close together and that the slope for F3 radiates upward while the slope for F2 radiates downward away from the loci. The data for /g/ in Fig. 5 shows that only two vowels, /ae/ and /a/, display the predicted compact pattern. For the front vowels, both F2 and F3 slopes radiate downward, while for the back vowels the loci are quite far apart in frequency.

The slopes for F1 are not shown in Fig. 5 but they can be observed separately for each vowel in Fig. 3. It has been claimed repeatedly that voiced stops have a rising F1 which is a manner cue to distinguish voiced stops from vowels, (Cooper et al., 1952; Stevens and House, 1956; Stevens and Blumstein, 1978). The measured F1 transitions for this speaker showed that five of the 24 average F1's were either flat or falling (/bi, di, gi, bI, gu/). Apparently, in natural speech, a rising F1 is not a necessary property of voiced stops spoken before high vowels.

The results from the measured loci and slopes of the formant transitions for this speaker can be summarized briefly. Only three invariant acoustic loci were found, both F2 and F3 for /d/, and F3 for /b/. F2 for /b/ had two

loci, and F2 and F3 for /g/ showed no clustering of loci across vowels. The predicted pattern of rising F2 and F3 formants for /b/ was not observed for all vowels in these data. Compact formant frequency transitions for /g/ were not observed for this subject for six of the eight vowels. F1 formant slopes did not rise for some stops with high vowels, and therefore, cannot be a necessary acoustic correlate of voiced stops in English.

Finally, voice onset time (VOT) was examined as a possible correlate of place of articulation. VOT was measured as the difference in milliseconds between the frame in which voicing begins according to the criteria mentioned in Section C above, and the first frame which contained the release burst. Since VOT seemed more or less invariant across vowels, statistical analysis of VOT was collapsed across all eight vowel contexts.

Insert Table 4 about here

Table 4 displays the results of these analyses in which VOT has been converted from frames to milliseconds. The average VOT values measured for this speaker showed the characteristic differences in VOT for different places of articulation, namely, that VOT is longer as place of articulation moves from the front of the mouth (labial) to the back (velar) (Lisker and Abramson, 1964; Zue, 1976).

Table 4. Confusion matrix produced by Discriminant Analysis for assigning stops before eight vowels on the basis of VOT measures alone.

		Predicted stop category		
Stop	(VOT ms)	b	d	g
b	(3)	98% (n=39)	2% (n=1)	0% (n=0)
d	(11)	5% (n=2)	70% (n=28	25% (n=10)
g	(20)	0% (n=0)	5% (n=2)	95% (n=38)

Statistical analysis showed that VOT alone was, in fact, highly effective in classifying place of articulation. A Discriminant Analysis was then used to produce a confusion matrix for stop classification. Discriminant Analysis was done pairwise for /bd/, /dg/ and /bg/ using VOT to sort the stops into categories. The resulting stop categories sorted each of the 120 CV's unambiguously into a single category except for one /ba/. Thus, it was possible to collapse the pairwise matrices into one matrix by labeling the disputed /b/ as /d/, not as /g/. These results are shown in Table 4. The overall percent correct classification was 88%. Most of the errors occurred in classifying /d/. Although these results are for a single subject speaking at a constant tempo, VOT was a very effective property for classifying place of articulation, and, moreover, it appears to be context invariant across different vowels.

E. Discussion

An extensive set of formant transition parameters was examined in this experiment as possible acoustic correlates of place of articulation. As expected, little evidence of context invariant cues for place of articulation was observed in these data. For the onset frequencies of the F1, F2 and F3 transitions, /b/ was statistically grouped separately from /d/ and /g/ averaged over all vowel contexts. This tendency, however, was by no means a reliable acoustic correlate of place as shown in Table 3

and Fig. 4. These results from onset frequency measurements are in agreement with the earlier findings of Lehiste and Peterson (1961) and Fant (1973).

The only parameter displaying context invariant effects was VOT, which is not usually considered a formant transition parameter. VOT can be interpreted as a formant transition parameter in the sense that it is a durational measure specifying when the onset of the voiced F1 formant energy occurs relative to the release burst. It should be noted that this definition is somewhat different than the "F1 cutback" strategy used to simulate VOT in synthetic stop consonants which might be more appropriately termed a formant transition parameter (Liberman, Delattre and Cooper, 1958). F1 cutback was defined in this early study as the difference between the onsets of the F1 and F2 formants in synthetic stops which did not contain bursts. The measurements made in the present experiment were between the burst and F1 in natural stops.

Using VOT alone, our results showed that place of articulation was correctly classified for 88% of the syllables examined. While this is statistically an impressive result, there are several reasons why VOT alone should not be considered a possible invariant cue to place of articulation in stops. First, there are limitations on the ability of the human perceptual system to discriminate between the VOT values obtained in this study. The average value of VOT measured for /b/ was 3 ms and for /d/ was 11

ms. Several studies, including Stevens and Klatt (1974) and Pisoni (1977), have shown that the auditory system cannot discriminate between these values of VOT (but also see Sachs and Grant, 1976). For /g/, the average value of VOT was 20 ms. The studies previously mentioned showed that 20 ms marks the boundary of discriminability between simultaneous and non-simultaneous onset categories. These findings suggest that many /g/'s, having VOT's longer than 20 ms, could be perceptually categorized as distinct from /b/ and /d/. However, the results of this study showed that 25% of the /d/'s had long enough VOT's to be misclassified as /g/'s. Thus, VOT cannot be considered as a sufficient or reliable cue to place of articulation. Furthermore, in this study, VOT was examined only in voiced consonants. Klatt (1975) and Zue (1976) have shown approximately a 15% overlap between the VOT measures for /g/ and /p/. We conclude, therefore, that while VOT may not be a reliable acoustic cue to place of articulation, it might well be a secondary cue, particularly for separating velars from other consonants. Although this potential role of VOT as a cue for velar place may be small, it does appear to be context invariant.

The concept of an invariant formant locus for F2 and F3 was not supported by the results of this experiment. The calculated formant loci values agree with those measured in natural speech by Öhman (1966) and Fant (1973), as well as with the electrical analogue study of Stevens and House

(1956). Reasonably invariant acoustic loci were observed for F2 and F3 for alveolar stops, but not for the bilabials or velars. Apparently, the concept of an invariant locus for place for natural stop consonants remains unsubstantiated.

All the acoustic parameters measured in Experiment 1 were also examined for evidence of distinctive place information when the vowel context was known. Only the onset frequencies of F2 and F3 appeared to provide context dependent information for distinguishing place. Distinct place of articulation categories were obtained from F2 onset frequencies for six of the eight vowels studied; /i/ and /a/ were the exceptions. On the other hand, distinct categories were obtained for F3 onset frequencies for just one vowel, /ae/. Thus, the well-known Haskins view that the direction and extent of the F2 formant transition is an important place cue is consistent with the results of this study for most of the vowels examined. Nonetheless, our results suggest that the onsets of the F2 and F3 formant frequencies cannot alone serve as <u>sufficient</u> correlates to distinguish place of articulation for all vowel contexts studied.

When the F2 <u>and</u> F3 onsets were combined in a two-dimensional, F2 X F3 space separately for each vowel, place of articulation categories were very effectively classified. This was the only result in Experiment 1 which clearly provided evidence for distinguishing place of

articulation on the basis of acoustic parameters derived from measuring only the formant transitions. However, it is by no means obvious that humans make use of the combined F2 X F3 onset frequency information in normal speech perception. The combined F2 X F3 onset parameters were only effective potential cues when vowel context was known. Thus, unless we assume (as Haskins researchers typically have) that the basic unit of speech perception is the syllable (or perhaps the phoneme dyad), such that vowel and consonant identification are dependent on one another, the information from the F2 X F3 onsets can only serve as secondary cues for place identification. The interdependence of consonant and vowel perception is examined in Experiment 3.

It is, however, necessary to qualify any inferences to human speech perception from the findings of this experiment. The frequency resolution of the linear prediction method used to analyze the formant transitions is on the order of 50 Hz for measurements of the first three formants (Monsen, 1981). This frequency resolution most likely exceeds that of the human auditory system, which has variable frequency resolution bandwidths known as critical-bands (Scharf, 1970). In the human auditory system, frequency resolution in the F2 and F3 regions appears to be 2 to 5 times poorer than the resolution provided by linear prediction analysis. Implications of these findings for generalizing the results obtained in

this experiment need to be considered. The human observer cannot determine the onset values of F2 and F3 as accurately as they were specified by the digital analysis used here. Thus, on Fig. 4 the frequency boundaries of the ellipses denoting distinctive stop categories which are very close together (e.g. /b/-/d/ for the high vowels) will be blurred in the auditory system such that the distinctions between stop categories are likely to be lost. The results of this measurement experiment make few claims whose validity requires the detailed frequency resolution provided by the linear prediction analysis. In such cases, however, we should be careful in our interpretations of these results for human speech perception.

A number of results from Experiment 1 can be interpreted as providing counterevidence to widely claimed generalizations concerning the nature of stop consonant formant transitions. Two of these are considered below. It has been claimed on the basis of extensive experiments with synthetic speech (Delattre et al., 1955), as well as acoustic models of the vocal tract (Fant, 1960) that the F1 formant transition rises for initial voiced stop consonant syllables, and moreover that a rising F1 is a necessary cue for perceiving voiced stops as obstruents. Since the average transitions measured for 21% of the syllables in Experiment 1 were flat or falling, this generalization is not valid.

Another generalization concerns the formant transition patterns for velar stops. Stevens (1972; Stevens and Blumstein, 1978) has given a specific interpretation of the distinctive feature of "compact" for the proximal onsets of the F2 and F3 formant transitions. However, the loci and slopes shown for /g/ on Fig. 4 provide counterevidence for this claim for most of the vowels studied here.

We conclude from the results of Experiment 1 that invariant correlates of place of articulation are unlikely to be observed in the acoustic parameters of formant transitions of initial stop consonants. Furthermore, in this study only one complex measure of formant transitions proved to be even a potential, context-dependent correlate for place, namely the combined onset frequencies of F2 and F3 in a F2 X F3 space. Although great care was taken to examine place cues for only one talker speaking under ideal laboratory conditions, these efforts were not uniformly successful over all eight vowel contexts studied. Therefore, although acoustic correlates for place of articulation are quite distinctive in the formant transitions for some vowel contexts, it appears that formant transition parameters cannot generally serve as acoustic correlates of place of articulation.

Support for these conclusions can be found in recent speech perception studies using natural CV syllables. Dorman et al. (1977) deleted the burst and aspiration from stop-vowel CV syllables so that only the voiced transition

plus vowel segments were presented to subjects for consonant identification. The syllables consisted of /b,d,g/ produced before nine English vowels spoken by two talkers. Subjects identified the consonant better than 80% correct for only a third of the syllables. In a similar study dealing with the identification of voiced transition plus vowel segments, Ohde and Sharf (1977) obtained near chance (39%) average correct consonant identification of /b,d,g/ with three vowels. Apparently, in natural speech, the acoustic cues for place of articulation are encoded in more complex ways than are revealed by the analyses carried out in this experiment on formant transitions. In the next experiment, we will examine the changing patterns of spectral information continuously from the release burst into the voiced formant transitions.

III. EXPERIMENT 2. PLACE OF ARTICULATION FEATURES IN RUNNING SPECTRA

A. Introduction

We were unable to find reliable context invariant acoustic correlates of place of articulation in the analysis of voiced formant transitions in Experiment 1. However, it was obvious from the spectral sections of the CV's that well-formed peaks were continuous from the voiceless to the voiced portions of the CV waveforms. That is, the burst and the formant transitions were not separable or discrete segments in this analysis, as they are when viewed spectrographically. Instead, prominent spectral peaks in the burst were continuous with peaks in the formant transitions. The recent findings of Searle and his colleagues (Searle et al.,1979; Searle et al., 1980) provided the inspiration to examine running spectral displays in order to represent the changes in spectral prominences from the release burst into the voiced formant transitions.

As mentioned in the Introduction (Chapter I), Stevens (e.g., Stevens, 1967; Stevens and Blumstein, 1978) and Fant (1960; 1973) have claimed that invariant acoustic information for specifying place of articulation in stops can be found in the burst release and early portions of the waveform. Stevens (Stevens and Blumstein, 1978; Blumstein and Stevens, 1979; Blumstein and Stevens, 1980) has argued

that this information can be captured in a single 26 ms integrated spectral section. The research undertaken in this experiment examines the initial portion of the acoustic waveform, as Stevens and Fant have suggested, but uses running spectral displays for the spectral representation. Linear prediction running spectra provide a good spectral representation for speech because there is fine resolution in the frequency by amplitude domain while the time dimension is simultaneously preserved. Continuous running spectral representations of speech are presumably a better model of the spectral information output from the peripheral auditory system than are Stevens' single integrated spectral sections. That is, the most basic property of auditory neural signals is that they vary directly with the time variations of the input acoustic signal. In fact, Schroeder, Atal and Hall (1979) have modeled the spectral processing properties of the ear in a manner quite analogous to our calculation of a running spectrum. In their model Schroeder et al. state:

> The inner ear ... performs a running short-time spectral analysis in which the frequency coordinate f is represented by a spatial coordinate x along the length of the basilar membrane. We approximate this process by short-time Fourier transformations over successive 20-ms time windows. (p. 1647).

On the other hand, the concept advanced by Stevens (Stevens and Blumstein, 1978; Stevens and Blumstein, 1980) that the auditory system detects and integrates spectral energy over approximately 26 ms is not well supported in

the psychoacoustic literature. It appears to be true, as Stevens and Blumstein have remarked (1978), that the auditory system responds to abrupt changes in energy at signal onset and offset in special ways (Zhukov, Zhukova, and Chistovich, 1974; Delgutte, 1980). However, evidence of spectral integration during rapid changes in spectral energy for a period of time as long as 26 ms has not been proposed in the psychoacoustic literature.

Figure 6 shows six typical running spectral displays of stops in different vowel contexts. The first frame effectively displays 5 ms of the burst release. Subsequent frames are offset at 5 ms intervals. Further details of how these displays were generated are provided in the Procedure section below.

Insert Figure 6 about here

As can be seen, these displays show continuous changes in spectral prominence from the release burst into the voiced formant transitions. The onset of voicing in these displays is indicated by the appearance of a well-defined F1 peak at low frequencies.

Visual examination of the CV's from Experiment 1 revealed a number of potentially <u>invariant</u> acoustic features or attributes that could be used to identify place of articulation in the running spectral displays. The results from earlier pilot studies (Kewley-Port, 1979(a);

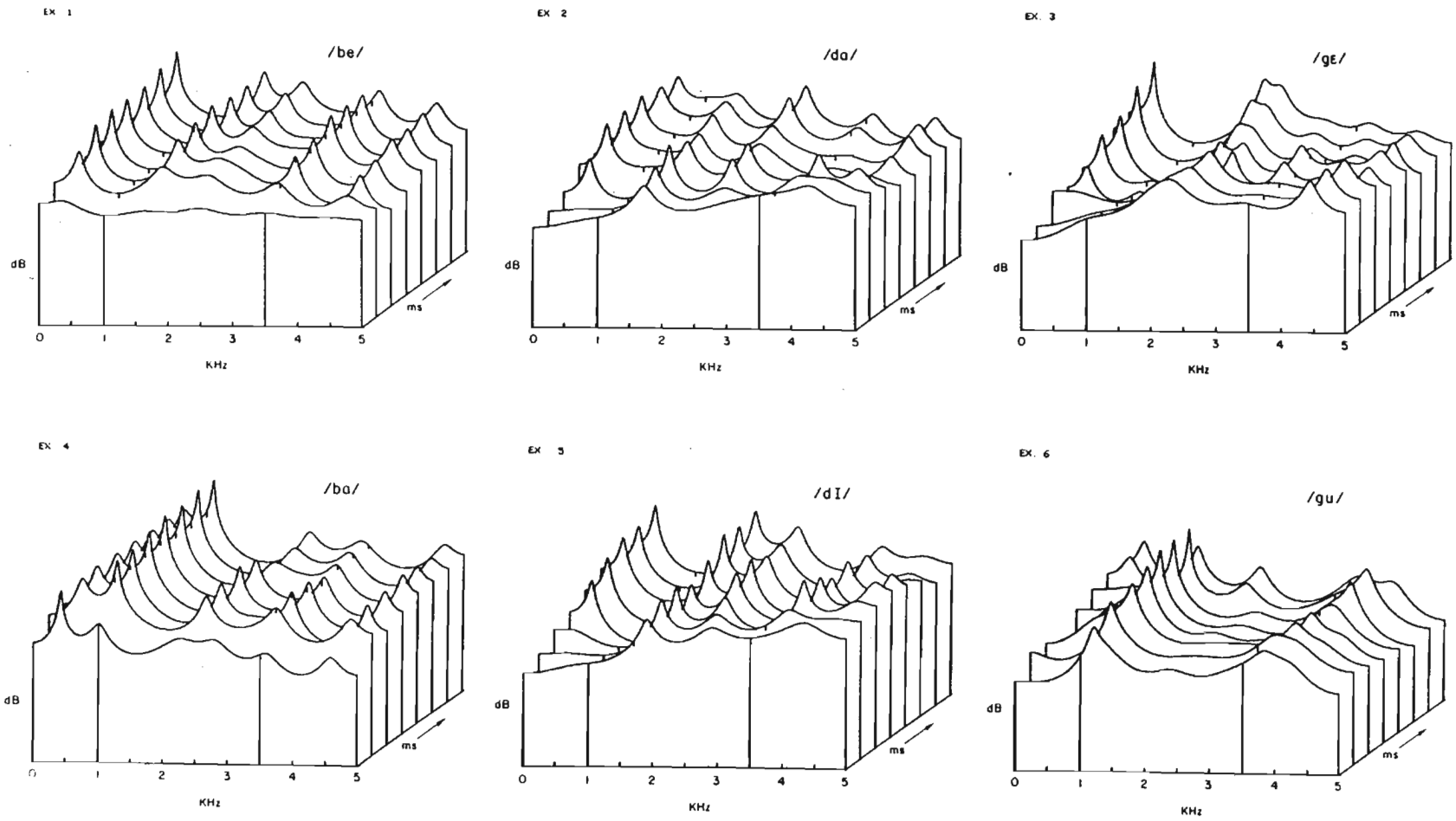


Figure 6. Running spectral displays for the first eight frames of six stop-vowel syllables. Each frame is offset by 5 ms.

1979(b)) of running spectra were also encouraging. The earlier feature definitions were carefully considered and redefined for the larger and more formal experiment presented here. The features that we developed were similar to acoustic features derived from the acoustic theory of speech production as proposed by Fant and to the features used by Stevens (Stevens, 1975; Stevens and Blumstein, 1978) in his research. In fact, as will be seen below, the features have strong roots in the distinctive feature theory of Jakobson, Fant and Halle (1952).

The features developed in Experiment 2 were specifically defined to identify place of articulation from the running spectral displays and were represented in terms of the following binary categories:

Feature 1: Tilt of the spectrum at burst onset. Tilt was estimated by visually fitting a straight line to the first frame between 0 and 3500 Hz. The feature categories were R=rising and F=flat or falling. Alveolar stops have been characterized as having rising burst spectra, and bilabial as having falling spectra. Velar bursts cannot be categorized as rising or falling, since burst energy shifts from high to low in English as vowels move from front to back. (Halle et al., 1957; Fant, 1960; Stevens, 1975; Stevens and Blumstein, 1978). The binary feature, Tilt of burst, is reminiscent of the Jakobson et al. distinction between "Grave" and "Acute."

Feature 2: Late onset of low frequency energy. Late onset was defined as the occurrence of high amplitude, low frequency peaks (F1 peaks) starting in the fourth frame of the display or later. Feature categories were L=late onset and N=no late onset. As mentioned in Experiment 1, the onset of the F1 peak is a measure of VOT and was very effective in classifying place for that one subject. The principle underlying the Late onset categories is that /g/ has a longer VOT than /b/ or /d/, usually greater than 20 ms. (see Lisker and Abramson, 1964; Fant, 1973; Zue, 1976).

Feature 3: Mid-frequency peaks extending over time. This feature was defined as the presence of a single, prominent peak between 1000 and 3500 Hz occurring for three or more frames, but not necessarily consecutive frames. The feature categories were Y=yes, peaks exist and N=no such peaks are present. These peaks arise from the resonant cavity in front of the velar constriction (Fant, 1960; Fant,1973). The acoustic prominence of this resonance has been called the "compact" spectrum associated with velar place by Jakobson et al. (1952), Fant (1960) and Stevens (1975).

After the feature categories are specified, place of articulation is assigned as /b/, /d/ or /g/ in accordance with the following assignment matrix:

Tilt of burst	late onset	Mid-freq. peaks	Assigned consonant
F	N	N	b
R	?	N	d
?	L*	Y	g

An entry of '?' in the assignment matrix means that either feature category may occur for that stop. The '*' by the feature L indicated that in ambiguous cases, the presence of L was sufficient to assign the stop g.

A pilot study using only the front vowels from Experiment 1 showed that observers could successfully judge the visual features and identify place in the running spectral displays (see Kewley-Port, 1979(a); 1979(b)). The purpose of the present experiment was to add more vowels and more speakers to the original data base. In addition, the features and assignment matrix were changed slightly from the original to the form presented above. Additional data were collected in this study to determine if the features and new assignment matrix could adequately describe the invariant visual properties of running spectra for identifying place of articulation in stop consonants.

B. Stimuli

Three talkers, two males (RP and TF) and one female (NL) produced the set of consonant-vowel syllables that were analyzed in this study. Syllables from the first talker, RP, were a subset of those analyzed in Experiment 1. They consisted of three repetitions each of /b,d,g/ paired with /i, I, e, ɛ, ae, a, o, u/. Examination of these syllables and the results of the pilot study were used to arrive at the definitions of the features and the assignment matrix presented above.

The data base was then expanded by adding an additional male talker, TF, and a female talker, NL, both of whom were phonetically naive. Their utterances were recorded in the same manner as those of RP. They produced the syllables /b,d,g/ paired with /i, e, a, o, u/ in the carrier phrase, "Teddy said CV." Three repetitions of each syllable were digitized for analysis from the middle of the ten lists recorded. The total number of syllables examined in Experiment 2 was 162, 72 from speaker RP, and 45 each from speakers TF and NL.

All syllables were analyzed using the SPECTRUM program to produce running spectral displays such as those shown in Fig. 6. The waveforms were edited and first-differenced (pre-emphasized). A Hamming window was then positioned so that the burst onset of the CV was in the center of the window. Linear prediction coefficients were calculated for

each window using the autocorrelation method of Markel and Gray (1976). Since this method requires at least two pitch pulses to fall within the analysis window, 20 ms windows were used for RP and NL, but TF, having a lower fundamental frequency, required a 25 ms window.

Spectral sections were then calculated at 5 ms intervals or frames. The Hamming window had been positioned in such a way that the first frame encompassing the burst had an effective duration of 5 ms. Therefore, the first frame can be said to display spectral energy from the release burst only. Very often the first several frames analyzed were voiceless, i.e., missing the F1 peak. According to Markel and Gray (1976), fewer linear prediction coefficients are needed to specify the spectrum adequately in voiceless than in voiced frames. Spectral sections calculated with fewer coefficients had smoother peaks, closer to the underlying fricative spectrum, than did the rippled peaks often produced by extra coefficients. Therefore, in this analysis, four fewer coefficients were used in analyzing the voiceless frames. Fourteen coefficients were used to calculate voiced frames for the males with five formant peaks; twelve coefficients were used for the female with four formant peaks.

A running spectral display was then plotted for the first eight frames -- or 40 ms -- for each CV. Using a Tektronics Model 4631 hard-copy unit, an 8 1/2 by 11 inch figure was produced for each of the 162 CV's. Figure 6

shows examples of six running spectral displays. All the CV's were randomized, coded by number, and placed in loose-leaf notebooks for examination by the judges in this experiment.

C. Judges

Three members of the laboratory who were not familiar with running spectral displays served as judges. Phonetically sophisticated judges were required for this task because the descriptions of the displays and phonetic features employed standard acoustic and phonetic terminology. Two of the judges (ACW and TDC) were graduate students in Psychology, and one was a post-doctoral fellow (SEK) with a Ph.D. in Speech and Hearing Sciences.

D. Procedure

The present experiment consisted of three parts; (1) training, (2) independent judging and (3) collaborative judging. The training session was used to acquaint the judges with the feature definitions and the assignment matrix to be used to identify place of articulation from running spectral displays. A typewritten page containing the feature definitions and assignment matrix for consonant identification as described earlier was given to each judge. (See Appendix C for a copy of the definition sheet.) A 20 minute training session was conducted by the experimenter with 15 examples of running spectra, none of

which were included in the 162 test displays. Figure 6 shows the six primary examples used in the training phase. Table 5 shows the correct feature responses and consonants on a facsimile of a response sheet.

Insert Table 5 about here

During the training phase, judges learned to judge each feature category for a display independently and write corresponding letters on the response sheets. Then judges were asked to assign a consonant according to the entries in the assignment matrix. It was noted that the assignment matrix did not include all possible combinations of features. The judges were told, however, that combinations not represented in the matrix would probably occur infrequently. If they occurred, judges were instructed to assign a consonant in whatever way they saw fit.

After training, the observers were asked to judge each of the 162 displays <u>independently</u>. Each judge scored half of the displays in the loose-leaf notebooks in two separate one-hour sessions on different days. The response sheets from the independent sessions were scored for correct consonant identification only. One or more incorrect responses occurred on 46 of the 162 displays. To resolve errors which may have resulted from careless judgements, a collaborative judging session was also arranged one week after the independent judging sessions were completed. The

Table 5. Correct responses for the training examples seen in Fig. 6.

Example	Features			
	Tilt of burst	Late onset	Mid-freq. peaks	Assigned consonant
1	F	N	N	b
2	R	N	N	d
3	R	L	Y	g
4	F	N	N	b
5	R	L	N	d
6	R	L	Y	g

three judges met together with the experimenter and were given the feature definition sheets and additional instructions for rescoring the 46 displays in which errors occured. Judges were instructed to write down whether they unanimously agreed or disagreed on a response set for a given display. When judges disagreed, they were asked to indicate in what way the features or matrix were ambiguous. The displays were judged in one single 1 1/2 hour session which was taperecorded for later analysis.

E. Results

In the collaborative judging, 20 of the 46 displays were unanimously assigned to the correct consonant, while 10 of the displays were unanimously identified incorrectly. The remaining 16 were judged to be ambiguous. To obtain an overall score for correct identification in Experiment 2, all unanimous assignments from the independent and the succeeding collaborative judgements were used in the final results. For the 16 ambiguous displays, the forced choice responses from the independent judging were used since consonants had not been assigned for ambiguous cases in the collaborative judging. Percent correct consonant assignment in the independent judging for these 16 ambiguous displays was 42%, which is slightly better than chance (33%).

Insert Table 6 about here

Table 6. Results of assigning consonants to running spectral displays in Experiment 2 using three judges.

Talker	No. errors	(N)	% Correct	% Correct by sex
RP	14	(216)	94	male = 92
TF	14	(135)	90	
NL	30	(135)	78	female = 78
Total	58	(486)	88	

Table 6 shows the overall results for consonant identification. Consonants were correctly identified 88% of the time from the running spectral displays. The errors were not uniformly distributed by talker. Specifically, there was a large difference in correct identification depending on the sex of the talker, 92% correct for male talkers, but only 78% correct for the female talker.

Insert Table 7 about here

The distribution of errors by consonant and vowel is shown in Table 7. Overall, /d/ was identified the best, 93%. The consonant /b/ was accurately identified in most contexts except in syllables containing the high vowels /i,e,u/. Most errors occurred for /b/ syllables produced by the female speaker. Analysis of the collaborative judging errors indicated that the tilt of the burst was ambiguous or slightly rising for bilabial stops before high vowels. /g/ was poorly identified primarily in the syllable /gi/ with 56% correct. Most of these errors were also contributed by the female speaker whose mid-frequency peaks occurred above the 3500 Hz limit imposed in the original feature definitions. Additional /g/ errors occurred primarily because the otherwise prominent mid-frequency peaks were not clearly "single" peaks in the display.

Table 7. Percent correct identification of consonant by vowel for all talkers. Note, vowels /I, ɛ , ae/ were produced by only one talker.

Vowel	(N)	b	d	g	Total
i	(81)	67	93	56	72
e	(81)	70	81	89	80
a	(81)	100	93	100	96
o	(81)	93	96	100	96
u	(81)	78	96	100	91
I	(27)	89	100	100	96
ɛ	(27)	100	100	67	89
ae	(27)	100	100	67	89
Total	(486)	84	93	87	88

Insert Table 8 about here

Table 8 displays the pattern of errors for the consonants identified in the running spectral displays. /b/ consonants were frequently mistaken as /d/'s, but not the other way around. Bilabials and velars were rarely mistaken for one another. /g/'s were frequently mistaken for /d/'s.

The feature categories assigned for each CV were analyzed to determine whether the judges had reliably and consistently categorized the features. Judges were said to disagree on feature assignment when they had not unanimously assigned the same feature categories. There were three features judged for each display. Mathematically, if features were assigned randomly by three judges, a score of 67% disagreement should be obtained. The results showed that when the judges correctly identified the consonant in the independent judging, only 2% feature disagreement occurred. In the collaborative judging of the other 46 displays, only 8% feature disagreement was observed. Therefore, an overall score of 5% feature disagreement was obtained from the three judges. These results indicate that it was relatively easy for the judges to categorize the features in the running spectral displays as specified in the feature definitions.

The specification of the original assignment matrix was based primarily on running spectra from a male speaker,

Table 8. Percentage of response errors obtained for a given consonant in running spectral displays.

	Assigned Consonant		
Displayed	b	d	g
b	-	14%	2%
d	2%	-	5%
g	0%	13%	-

RP. To check the validity of the feature matrix for assigning stops for all three talkers, the percentage of feature assignments made by all three judges in the independent judging was calculated. These results are given in Table 9, where percentages are entered in terms of the categories as they appeared on the feature definition sheet.

Insert Table 9 about here

It can be seen from the data in this table that the percent judgement of feature categories other than '?' was quite high, averaging 92%. The two entries of '?', which signified that either category might occur, were assigned equally to the categories. These results indicate that the original assignment matrix was appropriate for the features examined in this experiment. The slightly lower assignments of correct categories for Tilt of burst, (85% and 88%), and the presence of Mid-frequency peaks for /g/ (83%) suggest that some improvement of these feature definitions may be needed in future studies.

Not all possible permutations of the feature categories were listed in the assignment matrix. As a consequence, there were possible combinations of feature categories which led to ambiguous consonant assignments. Most of these occurred for the 46 displays in which a consonant error occurred and were dealt with in the

Table 9. Percent of feature assignments obtained in independent judging. They are listed according to the feature categories appearing in the feature matrix used for consonant assignment.

Consonant	Tilt of burst	Late onset	Mid-freq. Peaks
b	F=85	N=96	N=98
d	R=88	? N=59 ? L=41	N=96
g	? F=35 ? R=65	L=96	Y=83

analysis of the collaborative judging data. However, for the displays judged unanimously correct, only 1% of the responses resulted from ambiguous feature combinations. Thus, possible ambiguities in the assignment matrix were not a problem in this study.

Only ten of the displays (6%) were unanimously assigned the incorrect place of articulation. The incorrect assignments for these displays provide some insights into problems with the current feature definitions. Four of the ten displays were /b/'s before front vowels. Each burst tilt was rising such that the judges assigned d's to these displays. Three other displays contained the female talkers /gi/'s. The small mid-frequency compact peak for these /gi/'s lay at or above the 3500 Hz frequency limit used in the feature definitions. Judges also incorrectly assigned d's to these displays. We will suggest below slight modifications of the feature definitions which may result in correct place assignments for some of these seven displays. Analysis of the remaining three displays (2%), however, showed bizarre compact peaks for which no obvious rule change or explanation appeared helpful.

Some special attention should be given to the feature Late onset of F1. All three features in this experiment were defined as binary. Since only two binary features are necessary to specify the three consonants, this feature system is redundant. In the earlier pilot study it appeared that the feature of Late onset was so prominent in the

running spectral displays that it ought to play some role in distinguishing place of articulation among the stops, especially /g/ from /b,d/. In particular, since Mid-frequency peaks were sometimes difficult to identify for /g/, it was thought that in ambiguous cases, the additional category of Late onset of F1 might facilitate the correct identification of the stop feature. This was incorporated in the feature definition sheet using "L*" in the matrix for /g/ (see section A above). Post hoc analysis revealed, however, that the Late onset feature was actually used by the judges to disambiguate /g/ in only three cases (.6% of the judgements). Thus, the current feature definition system did not adequately capture the intended usefulness of Late onset, even though Late onset was present in 96% of the /g/'s.

As a consequence, the running spectra were reexamined to determine if an alternative definition for Late onset could be developed. The frame in which the onset of F1 occurred was determined by the experimenter for the 162 displays. In doing this we found that approximately 50% of the /g/ displays had F1 onset in frame 6 (30 ms) or later, compared to only one /d/ display and no /b/ displays. This observation suggests a new definition for Late onset, namely, that the category L refers to onset of F1 peaks in the sixth frame or later.

The assignment matrix should then be changed to:

Tilt of burst	Late onset	Mid-freq. peaks	Assigned consonant
F	N	N	b
R	N	N	d
?	L*	Y*	g

The * notation means that any occurrence of Late onset or Mid-frequency peaks requires that /g/ be the assigned consonant. Results from the independent judging were then reexamined to determine what effect the proposed changes would have on the correct consonant identification. However, none of the earlier correctly assigned /g/'s were affected by this reanalysis. Of the /g/'s incorrectly assigned, 6% more correct /g/ identifications would be obtained. No incorrectly assigned /b/'s or /d/'s were affected. However, 1.8% of the previously correct assignments of /b/ or /d/ were now incorrectly assigned to /g/. Therefore, the proposed change in the Late onset feature produced only a small overall improvement in consonant identification if incorporated in the analysis. However, this change represents a much better implementation of the concept which was intended for the Late onset of F1 feature, namely, that the presence of the

Late onset of F1 (i.e., a long VOT) is strongly associated with the consonant /g/. Such a feature can be useful in disambiguating /g/ in running spectral displays where Mid-frequency peaks may not be clearly present.

F. Discussion

The results of this study demonstrate that invariant features for identifying place of articulation in initial stop consonants are readily observable in continuous running spectral displays. This experiment was an initial attempt to establish the adequacy of the analysis procedures and feature descriptions for identifying place across a large number of vowel contexts and several talkers. The results showed that the features appeared to be reasonably invariant over vowel context, but were not generally invariant over sex of the talker. Stop consonants from the two male talkers were more accurately identified than those from the one female talker.

In the sections below, three theoretical issues regarding the use of running spectral analysis will be addressed. First, how does the present analysis relate to previous research describing the acoustic cues for place of articulation? Second, what is the relation between the linear prediction running spectra displayed here and representations of changing spectral information in the human auditory system? And third, what specific improvements can be made in the feature definitions and the analysis procedures used to identify place of articulation?

1. Acoustic cues for place of articulation

The features proposed in this chapter are conceptually closer to the acoustic cues for place proposed by Fant (1960; 1973) and Stevens (Stevens and Blumstein, 1978) than those proposed by Liberman and his colleagues at Haskins Laboratories (Liberman et al., 1954; Liberman et al., 1967; Liberman and Studdert-Kennedy, 1978). Liberman has stated that an essential property of human speech is that the acoustic cues for phonemes are encoded and overlapping in time. Thus neither the phoneme, nor the underlying articulatory gesture can be perceived directly in the speech signal. According to this account, there are no acoustic cues for place that are invariant over phonetic variables such as vowel context, talker, and rate of speaking. Fant and Stevens, on the other hand, have specified relational acoustic properties in the speech signal that directly reflect the underlying articulatory gestures. Thus, according to the principles of the acoustic theory of speech production, a general description of the acoustic cues for place in stop consonants is clearly specifiable. Furthermore, since the articulatory gesture for a given consonant is assumed to be relatively fixed, the corresponding acoustic cues for place should be invariant. The running spectral features identified in the present study appear to be exactly the type of relational, invariant features originally suggested by Fant and Stevens.

In several recent papers, Stevens and Blumstein, have proposed their own set of place of articulation cues in the form of invariant onset spectra (Stevens and Blumstein, 1978; Blumstein and Stevens, 1979). Their analysis differs from the present one in primarily one way; running spectral features incorporate the time dimension whereas Stevens and Blumstein onset spectra are basically static spectral "snap-shots" of the continuously changing speech signal. Hypotheses concerning the role of the acoustic cues in human speech perception are substantially different for the time-varying features proposed for the running spectral analysis here than for the static onset spectra proposed by Stevens and Blumstein. These differences are discussed below. It should be noted, however, that Blumstein and Stevens (1979, p. 1013) have acknowledged that the time dimension might have to be incorporated in their analysis. In fact, after our pilot work was completed (Kewley-Port, 1979a; 1979b), Blumstein and Stevens (1979, p. 1013) suggested that an analysis procedure similar to the running spectra of Searle et al. (1979) might be an improvement over their single-spectrum static template procedure.

At this point it may be useful to examine the temporal properties of the running spectral features. Up until now the specific role of timing for the feature Tilt of burst has not been discussed because the spectral section of the burst release was positioned in the first frame of the running spectral displays (see Fig. 6). How would the burst

spectrum be identified in running spectra that tracked spectra throughout the stop closure before the burst release? Acoustic features for locating the burst in running spectra have already been proposed by Stevens (Stevens, 1980; Stevens and Blumstein, 1980). The release burst can easily be detected when an abrupt change in energy occurs, followed immediately by a rapid change in the spectral distribution. Stevens proposed that these properties were the acoustic correlates associated with the distinctive features of 'abrupt' and 'consonantal.' Both these acoustic properties incorporate change over time and should be easily observed in running spectra.

Another temporal property of running spectra affecting the Tilt of burst feature is the integration time window. The analysis of any speech spectrum requires a time window over which to integrate the spectral energy. In the case of time varying spectral representations, the window and the speech signal slide along in time relative to one another, usually in such a way that the resulting spectra have somewhat overlapping time windows. For stop consonants, a long period of silence, containing perhaps low energy voicing during closure, precedes the burst. As the time window slides along, it will eventually include only the first 5 ms of the burst, as long as the window is greater than 5 ms wide. In linear prediction analysis, a time window greater than 5 ms can, of course, be specified for running spectra. In the auditory system, rapidly changing

spectral information must also be integrated over some short time window. Estimates of the temporal resolving power of the auditory system range from about 20 ms (Hirsch, 1959; Miller, Wier, Pastore, Kelly and Dooling, 1976) to about 2.5 ms (Patterson and Green, 1970). Thus, the assumption that is implicit within the running spectral analysis is that the short-term integration of spectral energy can be set within limits of 5 to 20 ms, values that are in accordance with psychophysical measurements in the human auditory system.

The fact that running spectral displays are good at capturing a spectral frame containing only the burst makes this analysis technique closely compatible with Fant's theories (1960; 1973). Fant (1960, sec. 2.63; 1973, p. 135-137) proposed that place of articulation information is contained in the first 10 to 30 ms after the burst release. He discussed both spectral and temporal properties of the burst release which can be captured very effectively in the running spectral analysis. The spectral properties are: burst spectra for labials are spread and emphasize low frequencies; burst spectra for alveolars are spread and emphasize high frequencies; burst spectra for velars are compact in the mid-frequency region. These spectral properties of the burst are incorporated in the category definitions of the Tilt of burst feature and the consonant feature matrix described earlier.

Fant's descriptions, however, also involve temporal aspects of the bursts. Labial and alveolar bursts are said to be between 5 and 10 ms in length which is shorter than the 20 to 30 ms velar bursts. Furthermore, velars are described as having a compact spectrum which lasts throughout the longer burst. This compact spectrum arises from a resonant pole produced in the cavity in front of the velar constriction. Fant pointed out that the velar release is relatively slow so that this resonance is sustained for approximately 30 ms. A distinctive property of velars is that only slow changes in spectral energy are observed over this interval compared to the rapid changes observed for labials and alveolars. The running spectral analysis as proposed here also captures these temporal properties of the release bursts. The 5 ms burst frame is quite prominent in this analysis and clearly displays the tilt of the bursts. As the window slides along, for bilabials and alveolars, transient energy is encountered and a rapid change in spectra can be observed (see Fig. 6). For velars, however, successive spectra following the burst show little change in the prominent mid-frequency spectra.

These properties of the running spectral analysis can be contrasted with the static template analysis of Stevens and Blumstein (Stevens and Blumstein, 1978; Blumstein and Stevens, 1979). It is important to note here that Stevens and Blumsteins' theoretical framework regarding the spectral properties of the burst is in complete accordance

with Fant (Stevens and Blumstein, 1978). Unfortunately, however, their analysis based only on onset spectra obscures the important spectral differences that exist for the labial and alveolar bursts by integrating energy over a 26 ms time window. A 26 ms window will always include some transitional information about the vowel (voiced or voiceless) along with the energy in the burst. Furthermore, a single, fixed integration window cannot account for the differences in the temporal properties of the release burst as described by Fant. The rapidly changing spectra following the burst for bilabials and alveolars cannot be observed in only a single 26 ms onset spectra. But more importantly, the velar property of a slowly varying compact spectrum extending in time cannot be represented in only a single onset spectrum having no temporal dimension. We conclude, therefore, that the moving time window of the running spectral analysis used in the present study is a substantially better analysis technique than the Stevens and Blumstein static onset spectrum analysis using a fixed 26 ms window.

Another aspect of the Tilt of burst feature is the implementation of the concept of "rising" or "falling" energy as the slope of a visual regression line. The results of the present study demonstrated that the flat versus rising Tilt of burst contrast was successful in correctly classifying the stops. However, Table 8 showed that while 14% of /b/'s were classified as /d/'s, only 2%

of /d/'s were classified as /b/'s. This implies that the strict definition of flat versus rising Tilt produced a disproportionate number of /b/ errors. Several suggestions can be made for improving the definition of the Tilt feature. First, if linear prediction spectra are used, the actual slope of the regression line through the burst frame could be calculated for a large number of talkers. The distributions of the slopes could then be examined for /b/'s versus /d/'s, and an empirically defined slope for the rising versus falling categories could be determined. Another possibility is to employ a different measure of Tilt. Searle et al. (1979), attempted to capture the rising versus falling aspect of burst spectra by weighing spectral energy above and below 1700 Hz. They did not provide information in their report on the success of this measure in distinguishing /b/'s and /d/'s, but something like this measure could be explored for implementation with linear prediction running spectra. Two other aspects of the current definitions for Tilt of burst will be discussed later, namely, the adequacy of linear prediction spectra compared to other possible spectral representations of speech, and the differences in Tilt associated with differences in the vocal tract size of the speaker.

The feature Late onset of F1 represented an attempt to use differences in VOT as an acoustic cue for place of articulation. The results showed that the definition of this feature used in the current study did not contribute

to the correct classification of velar stops, and a new feature definition (with accompanying assignment matrix) was proposed. The new definition incorporated a specific place distinction discussed by Fant (1973, p. 136). Fant suggested that delayed voicing onset is a "secondary correlate to the place of articulation" for velars. Based on the results of the present analysis, and Fant's earlier suggestion, the Late onset of F1 feature should be treated as a possible secondary cue to place in future research. If Late onset is indeed a secondary cue to place, it might be worthwhile to determine if any empirical evidence could be found for the use of this feature in the perception of place of articulation. Several studies have already demonstrated that identification of place of articulation changes when the VOT of the stimuli is manipulated (Sawusch and Pisoni, 1974; Miller, 1977; Oden and Massaro, 1978). All three of these studies, however, used synthetic stop-vowel syllables without release bursts. Therefore, a more careful synthesis study using velar syllables with bursts containing the important Mid-frequency peaks feature is needed to demonstrate the specific role of the Late onset feature in perception of place of articulation.

The presence of mid-frequency spectral peaks in a stop burst has long been regarded as a primary cue to velar place of articulation, at least by researchers investigating the acoustic correlates of distinctive features (Jakobson et al. 1952; Halle et al., 1957; Fant,

1960; Stevens and Blumstein, 1978). We mentioned earlier in the discussion of the Tilt of burst feature that the running spectral analysis was very successful in capturing both the spectral and the temporal properties of the compact burst described by Fant (1960; 1973). Spectrally, a single prominent peak in the mid-frequency region was usually observed in the linear prediction spectra; temporally, these peaks appeared in successive frames representing the underlying articulatory gesture of the relatively slow velar release.

These findings may be compared to the Blumstein and Stevens (1979) onset spectra analysis with regard to identifying velar place. Blumstein and Stevens have specifically acknowledged in their recent publications that the compact spectra for velars must persist in time for listeners to identify velars correctly (1979, p. 1002; 1980, p. 652). Nevertheless, their single onset spectra cannot in principle represent this acoustic information adequately. The velar template constructed by Blumstein and Stevens (1979) is essentially a peak detector. However, in carrying out their analysis, they discovered that simple integration of the first 26 ms of spectral energy produced numerous spectra containing prominent peaks which were not velars. To be precise, 27% of the alveolar consonants had spectral peaks near the F2 locus at 1800 Hz. This observation prompted Stevens and Blumstein to ad hoc modification of their original diffuse rising template so

as to exclude peaks occurring around 1800 Hz (1979, p. 1005). Furthermore, alveolar peaks said to arise from subglottal resonances also occurred in the 800-1600 Hz region. These peaks were also arbitrarily excluded from the diffuse-rising template (1979, p. 1005). Little information was provided in Blumstein and Stevens' report concerning the possibility that the velar template incorrectly identified these peaked alveolar spectra as velars. However, from Table I in Blumstein and Stevens (1979) it can be seen that 17% of the /d/'s and 12% of the /t/'s were incorrectly identified as velars. Moreover, additional problem for the compact template was the presence of double peaks. Blumstein and Stevens (1979, p. 1006) apparently treated two spectral peaks separated by less than 500 Hz as a single peak, although no information was provided concerning how frequently such peaks occurred. Thus, it appears that attempts to locate the spectral feature compact as a simple peak in a single, 26 ms integrated spectrum have given rise to numerous exceptions and to the postulation of ad hoc decision rules. Similar problems did not arise in our analysis of running spectral displays, and we suspect that they did for Stevens and Blumstein because the temporal dimension of the speech signal was eliminated in their static onset spectra analysis.

2. Auditory representations of speech signals

The results from this study have demonstrated that human subjects can identify place of articulation from visual features displayed in running spectra. While such findings were reliable and consistent across three judges for a large number of natural speech tokens, there still remains the question of the relation between linear prediction running spectra and spectral processing of speech in the human auditory system. Specifically, it is important to know if the features used in the visual experiment are as robust in auditory spectral representations as they were in linear prediction spectra.

Several investigators have constructed auditory processing models to produce spectral representations of speech signals. In some cases, the output of these models has been examined for the presence of spectral cues for speech recognition. In particular, as noted earlier, the research of Searle and his colleagues (Searle et al., 1979; Searle et al., 1980) using 1/3 octave filters to search for place and voicing cues in stops inspired the development of the running spectral displays for the present study. Although Searle et al. explicitly advocated the development and use of auditory processing techniques for speech, the specific processor they implemented in their study did not adequately model known psychophysical properties of the human auditory system. In fact, they chose as the basis of

their speech processor a standard, commercial set of 1/3 octave filters. Design characteristics of 1/3 octave filters, in fact, do not model properties of the human auditory system very well, but were developed to meet a set of engineering standards for commercial filters (American National Standards Institute ANSI S1.11-1966 class III). The popularity of 1/3 octave filters for speech processing (e.g., Klein, Plomp and Pols, 1970; Schouten and Pols, 1979) derives from their general availability and speed (i.e., analogue processing). As we shall see below, however, their filter characteristics are only a gross approximation to the filtering properties of the human auditory system.

Before evaluating the success of these recent proposals it will be useful to review the essential spectral properties of running spectral displays with reference both to known properties of mammalian auditory systems and to other auditory processing models currently employed in speech analysis. In terms of the spectral sections, we will briefly discuss the characteristics of the analyzing filters (e.g., linear prediction, FFT, 1/3 octave filters), the representation of the frequency dimension (e.g., linear, log, bark), and the representation of the amplitude dimension (e.g., dB or sones). In addition, we will also discuss how the temporal dimension has been used to represent the changes in spectral energy over time.

Differences in design characteristics of the filters for processing speech signals can produce quite different spectral displays. Based on psychophysical measures, the frequency resolution in the human auditory system has been described in terms of a set of critical-bands (Scharf, 1970). A critical-band analysis corresponds roughly to the frequency analysis of a set of bandpass filters whose bandwidth is constant (about 100 Hz) below 500 Hz, and then becomes successively broader as frequency increases above 500 Hz. Bandwidth, however, is only one property of a bandpass filter. Two other properties less frequently discussed are the shape of the filter itself (e.g., rectangular versus Hamming) and the slopes of the skirts of the filter. (See Klatt 1976 and 1979 for a comprehensive list of appropriate design characteristics for critical-band filters for speech processing.) The discussion that follows emphasizes differences in bandwidth since it has received the most attention in the psychophysical literature.

Research has produced two sets of estimates for the bandwidth of critical-band filters, one about one-half as wide as the other (see Sever and Small, 1979). This range varies approximately from .1 to .18 times the center frequency of the filter. One-third octave filters have bandwidths approximately .23 times the center frequency. Thus the bandwidths of 1/3 octave filters are considerably broader than estimates of the critical-bandwidths derived

from psychophysical data. This means that 1/3 octave filters provide poorer frequency resolution for high frequencies than does the human auditory system. (See Flanagan and Christensen (1980) for a demonstration of high frequency differences between 1/3 octave filters and 1/6 octave filters.) Thus 1/3 octave filtering of speech signals probably represents a lower limit on the poorest frequency resolution that the human auditory system might display.

On the other hand, frequency resolution as measured in discharge patterns of auditory-nerve fibers in cats can be extremely accurate under optimal signal conditions. The analyses carried out recently by Delgutte (1980), and Young and Sachs (1979; Sachs and Young, 1980) have shown considerable accuracy in determining formant frequencies of synthetic vowels under certain conditions. It is not known if such precise formant frequency information is available under normal listening conditions for human speech recognition. However, the linear prediction spectra used in the present study provide a precise frequency resolution of speech spectra using a constant 19.5 Hz bandwidth. These linear prediction spectra when displayed on a log-frequency axis appear roughly similar in frequency resolution to many of the vowel-formant spectral figures reported by Young and Sachs (1979). Thus, linear prediction spectra may be considered as providing an upper limit on the best possible frequency resolution the human auditory system might have.

Several investigators have implemented critical-band filtering models for speech signals in which the bandwidths lie between the extremes mentioned above. Bladon and Lindblom (1979) used auditory filters one Bark wide for analyzing vowel formant patterns. (The Bark scale was developed by Zwicker and Feldtkeller (1967) to convert frequency to standard critical-band measurements that approximate about 1/4 octave intervals above 1 kHz.) Carlson and Granstrom (1980) have also used the Bark auditory filters, and compared them with 200 Hz wide FFT filters for specifying differences in vowel formant spectra. Zwicker, Terhardt and Paulus (1979) have also developed a set of 24 filters using Bark bandwidths for automatic word recognition. And Klatt (1979) has described an auditory filter bank for processing speech as input to a phonetic (Scriber) or word (Lafs) recognition system using filter bandwidths of approximately a quarter of an octave. (These values are similar to Barks.) Flanagan (1980; Flanagan and Christensen, 1980) has designed and implemented an auditory filter vocoder system, which tested bandwidth values varying from full octave (below 1000 Hz.), to 1/6 octave. Thus, there is considerable exploratory research underway at the present time to analyze speech signals using auditory filtering with approximately 1/4 octave (Bark) bandwidths.

With spectrally analyzed speech, several possible representations of the frequency by amplitude dimensions

can be chosen. Linear prediction spectra are typically represented on a linear frequency scale. In the auditory system, however, frequency on the basilar membrane is equally distributed in approximately Bark intervals (Schroeder et al., 1979), which is often approximated by a simple log-frequency scale. Thus for research employing auditory filters, Bladon and Lindblom (1979) and Carlson and Granstrom (1980) have used Bark frequency scales, while Klatt (1979) has used a modified log scale (technical Mel). Several researchers who have used auditory filters have specified loudness according to equal-loudness contours in sones (Bladon and Lindblom, 1979; Carlson and Granstrom, 1980).

The last property of running spectral displays to be discussed is the representation of time. We know that the auditory system can closely track time variations in waveforms in terms of synchrony of discharge firings with the input signal (Kiang, 1980). Apparently, the important acoustic distinctions in speech vary much more slowly than the temporal processing capabilities of the ear. Therefore, the limits of the representation of the time dimension for speech spectra should be set according to the observed rates of change in the speech signal. For speech, this limit would be placed somewhere between 1 ms and 20 ms. Searle et al. (1979) originally used a 1.6 ms time frame for running spectra. This time frame seemed to present too much detail, so they used averaged time frames of 8 ms in

their feature analysis. The running spectra in Experiment 2 were 5 ms apart, while Klatt's (1979) spectra were 10 ms apart. Thus the time intervals between spectra currently employed by different investigators are in the 5 to 10 ms range.

One other aspect of the temporal dimension should be mentioned. The time dimension in running spectra can be represented differently in analogue versus digital processing systems. The ear more closely resembles an analogue filter bank. As Searle (Searle et al., 1979) has pointed out, in analogue systems, the time delay encountered in analyzing a given spectral frequency is an inverse function of frequency. Thus high frequency spectral information is available before low frequency information. This property can be seen in the response patterns of auditory-nerve fibers (Delgutte, 1980). Differences in time delays in Delgutte's study were 4 ms between the .22 kHz unit and the 4.31 kHz unit. Searle has suggested that these time delays may provide useful information for speech processing, for example in rapid detection of burst onsets. On the other hand, digital speech processing techniques, such as linear prediction or FFT, do not usually represent these delays in running spectra. However, in running spectral displays with frame intervals of about 5 ms, little difference could be observed in displays generated by analogue and digital processing systems.

3. Running spectral features in auditory filter displays

The question originally examined in this experiment was whether running spectral features would be prominently displayed in auditory filter representations of running spectral displays. To explore this problem further, we decided to reexamine informally the acoustic stimuli with two auditory filter representations. SPECTRUM was modified so that a set of programmable, moving average filters could be applied to the previously computed spectral sections. Two properties of the filters were adjustable: bandwidth and the slopes of the skirts. Fixed properties included symmetry in a log-frequency space, and a flat top giving an overall trapezoidal shape. These fixed filter properties were choosen because they matched Searle et al.'s one-third octave filter specifications, although they are admittedly a crude approximation to auditory filter shapes (see Patterson, 1974). The filters were moved at short frequency intervals in order to produce smooth spectra. Pilot work showed that for the spectra used in Experiment 2, the convolution of the programmable filters with the smoothed linear prediction spectra produced filtered spectra almost identical to those produced by convolution with the equivalent 200 point FFT. Thus, the smoothed linear prediction spectra were used as input to the programmable filters in this analysis.

The programmable filters were selected as follows. One-third octave filters were chosen because they have frequently been used in speech processing (particularly by Searle et al., 1979 and 1980) and represent a frequency resolution which is probably poorer that that of the human auditory system. The 1/3 octave filters were digitally defined according to the ANSI S1.11-1966 class III standard, with a bandwidth constant of .23 times the center frequency, and skirts having a 50 dB/octave roll-off. The other filters, although similar to the auditory filters of Bladon and Lindblom (1979) and Klatt (1979), were patterned more closely after the narrower filters proposed by Patterson (1974) and the 1/6 octave filters of Flanagan and Christensen (1980). The bandwidth constant was .13 and the skirts had a 75 dB/octave roll-off. Below 400 Hz, regardless of the bandwidth constant, the bandwidth was fixed at 95 Hz in keeping with standard critical-band measurements (Scharf, 1970).

The running spectral display was also altered for spectra processed by the auditory filters by implementing a log-frequency scale. Amplitude was still displayed in dB, and the time frame rate was kept at 5 ms. With these changes, the displays previously examined could be redisplayed using 1/3 octave filters or auditory filters (of the Patterson type). Figure 7 compares three linear prediction running spectra with auditory filters displays. Figure 8 compares three different linear prediction running spectra with 1/3 octave filter displays.

Insert Figures 7 and 8 about here

We reexamined the displays computed earlier using both types of filters to determine to what extent the three visual features used to identify place were still present. Moreover, we were interested in determining how such filtering would potentially alter our earlier feature descriptions. Approximately half of the 116 displays which had been correctly identified by all three judges in the independent judging were examined first. Then all 46 displays from the collaborative judging were redisplayed and examined visually by the experimenter.

An examination of Figures 7 and 8 reveals that an auditory filter representation changes the frequency space in three important ways. First, the low frequency region of F1 is more prominently displayed. Second, filtering alters the spectral tilt of each spectral section. Because bandwidths are broader at higher frequencies, more energy is averaged into the high frequency filters causing an upward spectral tilt. Thus, both auditory filters and 1/3 octave filters result in a non-linear transformation of spectral tilt which emphasizes high frequency energy in comparison to the linear prediction spectra. Finally, the spectral peaks move toward higher frequencies because the filters are symmetrical in log-frequency space, which means that they include more high-frequency energy than

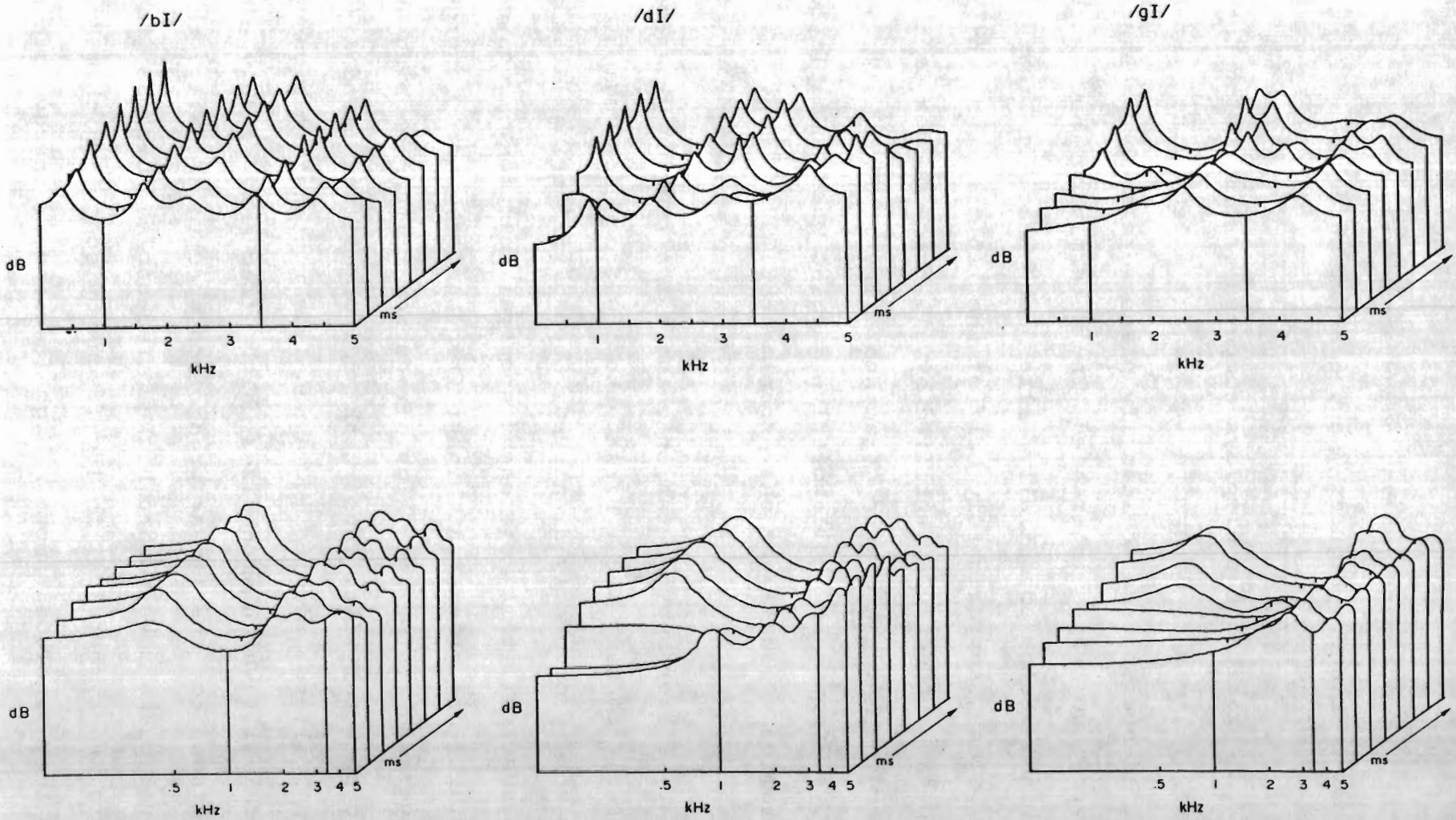


Figure 7. Comparison of running spectral displays produced by either linear prediction analysis (top), or smoothed by auditory filtering of the Patterson type (bottom) for three stop-vowel syllables.

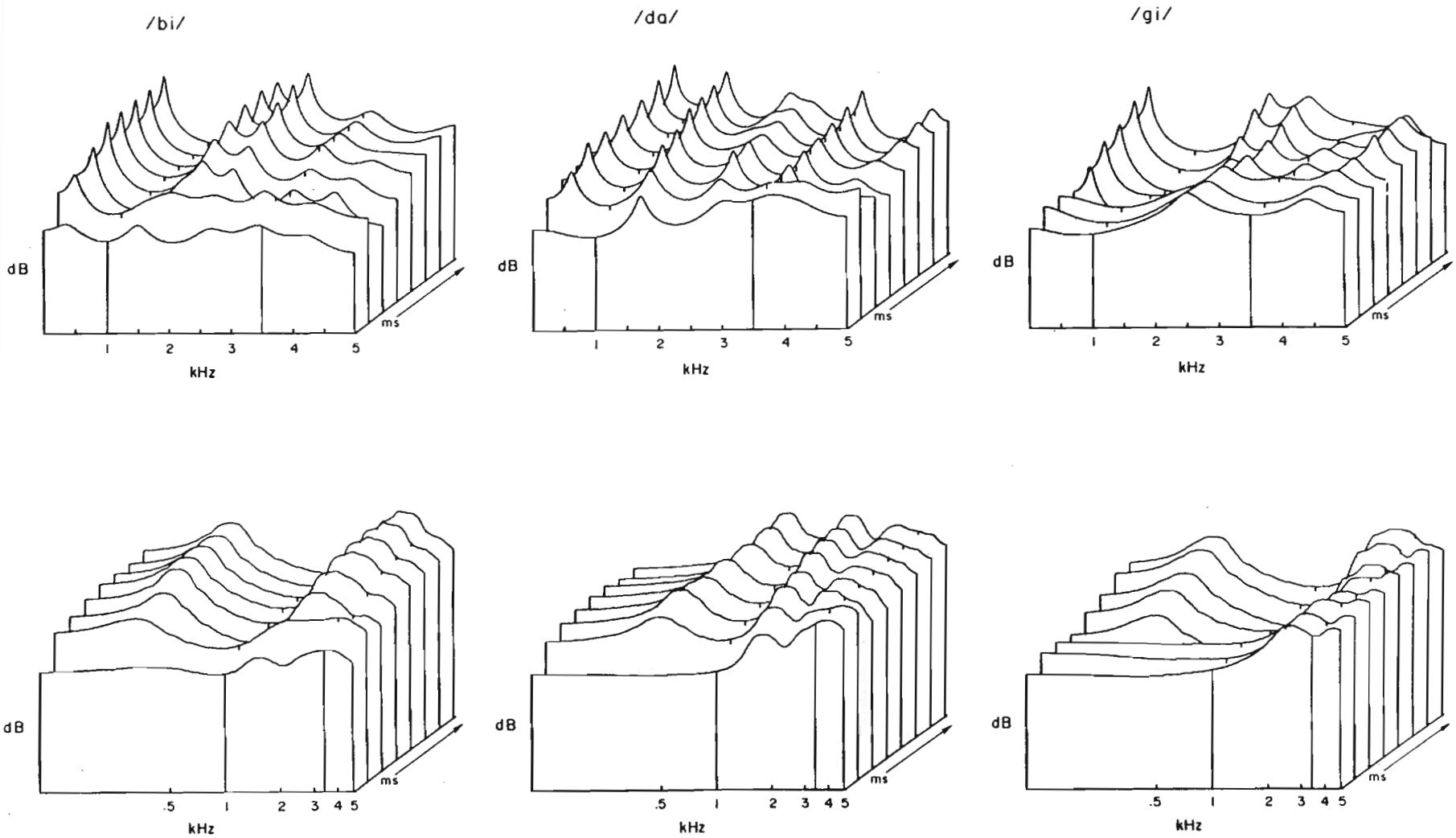


Figure 8. Comparison of running spectral displays produced by either linear prediction analysis (top), or smoothed by 1/3 octave filters (bottom) for three stop-vowel syllables.

low-frequency energy in a linear frequency domain. Klatt (1979) has recently implemented Patterson's (1974) hypothesis that auditory filters are symmetrical in linear frequency, but the log-frequency symmetry has been more commonly used. As a result of this shift, the feature definitions referring to the 3500 Hz marks shown on the displays should probably be altered to approximately 4000 Hz in the following discussion.

The results of the informal examination of the auditory filter representations of running spectra may now be compared to the earlier feature definitions based on linear prediction spectra. The Tilt of burst category definitions are altered in a similar way using either 1/3 octave or auditory filters due to the non-linear, high-frequency emphasis. The categories would be strongly rising for /d/ versus moderately rising for /b/. Looking at running spectra on which an error occurred such that a /b/ had been misclassified as /d/, the filtering appeared to disambiguate these cases specifically because of the non-linear emphasis on high-frequencies. However, the current auditory filter representations did not incorporate a transformation of the amplitude dimension from dB to sones. In sones, the relative amplitudes of the high frequencies output from auditory filters, would be reduced (see Bladon and Lindblom, 1979), which could compensate for the high-frequency emphasis of the filter. Thus the final effects of auditory filtering on the Tilt of burst feature

are yet to be explored. However, on the basis of our informal observations of the 1/3 octave and auditory filter displays, both filter representations of the Tilt of burst categories appeared to be as successful as the earlier linear prediction displays.

The definition of the Late onset feature is not altered significantly by filter representations of running spectra. The F1 peak, as shown in Figures 7 and 8, is described as a broad, well-formed low-frequency peak. Otherwise, no other aspects of this feature appeared to change for either 1/3 octave or auditory filters.

The visual display of the Mid-frequency peak feature changed quite a bit under the two types of filtering. Single prominent peaks were still readily observable for /g/'s, but they were narrower, had more "ripple", and appeared in more spectral sections (see Fig. 7). A significant problem occurred for the 1/3 octave filters which did not occur for the auditory filters. Many /b/'s analyzed by 1/3 octave filters acquired prominent mid-frequency peaks like those shown for /bi/ on Fig. 8. These peaks would cause them to be misclassified as /g/'s. For these same /b/ stimuli, auditory filtering did not produce mid-frequency peaks. Auditory filtering appeared to disambiguate many /g/ displays which had previously been misclassified either because two smaller peaks were superimposed on the prominent peak, or because the prominent peaks occurred on fewer than three spectra. Thus

the results from this analysis indicates that mid-frequency peaks for /g/'s are more salient in running spectral displays using auditory filtering than in those using linear prediction filtering. However, the 1/3 octave filter representations would cause many more /b/-/g/ confusions.

In summary, the present study demonstrated that three acoustic features could be used to identify accurately place of articulation from linear prediction running spectra prepared for visual inspection. However, linear prediction spectra provide a much finer spectral resolution than the human auditory system carries out. Therefore, two other spectral representations were used to construct running spectral displays. In the 1/3 octave filter representation, all features were preserved in the visual displays, but, unfortunately, too many mid-frequency peaks were erroneously produced for /b/'s. Thus, it appears that 1/3 octave filtering may eliminate some spectral properties that are important for speech analysis. These filters generate spectra having a poorer frequency resolution than that of the human auditory system. The other filter representation was representative of a class of auditory filters currently used by other investigators. These auditory filters have characteristics reflecting known psychophysical properties of the auditory system. The original features proposed in this study were all displayed robustly in auditory filter running spectra, and, in some cases, appeared to categorize place of articulation more

successfully than did the linear prediction running spectra. Thus, it appears that our three features can be used to accurately identify place of articulation in initial stop syllables displayed in two different time-varying, spectral representations of speech, namely linear prediction and auditory filtering running spectra.

4. Improvements in running spectral analysis

Based on the results of this study, and on the preceding discussion, it is now possible to recommend specific changes in both the definitions of the features and the analysis procedures which can be used in extensions of this research. Special attention will be given to the problem of vocal-tract normalization, i.e., accounting for the differences obtained between the male and female talkers in this study.

In terms of analysis techniques which compute the running spectra, further exploration of the auditory filter representations should be carried out. Since the human auditory system is the best consonant processor we know, implementation of analogous spectral processing capabilities may make the task of locating reliable acoustic cues for speech easier. This appeared to be the case for the auditory filters examined in this study.

Turning to the individual feature definitions, we begin with the feature of Mid-frequency peaks extending in time. The frequency range of 1000 Hz to 3500 Hz used in the

Mid-frequency peaks definition was determined from the earlier pilot study using male speaker RP. The results of the present experiment demonstrated that this range was unsuitable for the female talker because all potential peaks for /gi/ fell outside this range. Fant (1960; 1973) has already provided an explanation of this problem. In the 1973 paper he presented a table showing which formant is associated with the compact resonance peaks observed for /g/. For /i/ and /e/, the compact peak is associated with both F3 and F4. That is, for high front vowels, the more palatal constriction for /g/ produces a short vocal-tract resonance cavity. The formant peak of this cavity is high and close to F3 and F4. Thus, in order to capture a prominent peak for /g/, the mid-frequency range must be placed higher than F4 by approximately 500 Hz. The frequency range for F4 is, of course, dependent on a talkers' vocal-tract size and that will have to be considered in future analyses.

Likewise, the lower value of the mid-frequency range was originally set at 1000 Hz based on talker RP's velar peaks. The lowest peaks for velars occur before the vowels /o/ and /u/ which are associated with the talker's F2 resonance (see /gu/ on Fig. 6). Although the prominent peaks for /o/ and /u/ are continuous with F2, they fall from a higher frequency in the burst into the steady-state F2 for the vowel. Based on these observations it appears that a simple rule to account for vocal-tract size can be

implemented in the definition of the frequency range for Mid-frequency peaks which can solve this problem. The lower frequency limit should be placed at the lower frequency of a talker's F2 for /o/ or /u/, and the upper limit should be placed 500 Hz higher than a talkers' F4. These values are for linear prediction spectra. If auditory filtering is used, the lower limit would not change, but the upper limit should be set about 1000 Hz higher than the talkers' F4 because of the spectral averaging of the higher frequencies.

Next consider the Tilt of burst feature. It has previously been noted that more /b/ errors occurred for the high vowels than the low vowels, and more errors occurred for the female than the male talkers. That is, rising spectral tilts were sometimes obtained in the 5 ms, burst-only spectra for /b/, apparently due to high frequency energy associated with certain vowel contexts. Earlier in the discussion it was suggested that this problem might be solved by changing slightly the definition of rising tilt to be more prominently rising for /d/'s. It was also noted that with auditory filtering a non-linear emphasis of high frequencies was observed which had the beneficial effect of sorting more clearly /b/'s from /d/'s, particularly for the more ambiguous high vowel cases. We should note, however, that the Tilt of burst definition also includes an upper frequency limit, previously set to 3500 Hz. This limit was imposed because the burst-only

spectra for /d/ are not in fact "diffuse-rising" up to 5000 Hz as Stevens and Blumstein (1978) have suggested. Rather, for /d/ before back vowels, a vowel dependent peak in the spectra occurs for males at about 3000 Hz so that the spectra falls above this frequency. Both Zue (1976) and Klatt (1980) have previously reported this spectral property. This peak is vowel dependent and therefore varies with vocal-tract size. Thus, the upper frequency limit for the Mid-frequency peaks and Tilt of burst features must take into account vocal-tract size, and probably can be set to the same value using the previously suggested rule. Note that the problems in the feature analysis just discussed arose from an interaction of vowel context effects and differences in vocal-tract size. The specific solution proposed here is that by properly accounting for differences in vocal-tract size in the feature definitions, the features will automatically specify place of articulation independent of the vowel context.

Finally, the feature of Late onset of F1, as previously discussed, should be considered as a secondary feature for separating velars with long VOT's from bilabials and alveolars. Suggestions for a new definition of this feature and the resulting matrix were previously given at the end of the Results section.

The adequacy of these three features for specifying place of articulation in Experiment 2 was determined using human observers who examined visual displays of running

spectra. The presence of the three features in the running spectra could, in principle, be determined algorithmically by computer. However, since Experiment 2 was an initial study of these features, human observers were used to obtain feedback about possible ambiguities in the definitions or procedures in the collaborative judging session. In future experimentation with these features, a machine implementation of the feature definitions will certainly be included.

5. Conclusions

In conclusion, we propose three time-varying, relational acoustic features as a principled solution to the problem of invariant acoustic cues for place of articulation in initial stop consonants. These features are clearly observable in visual representations of running spectral displays of naturally produced CV syllables. The present study evaluated this proposed analysis for voiced stops before a large number of vowels produced by three talkers. From our results, it appears that these features are invariant over vowel context, and with the addition of two simple rules will be invariant over vocal-tract size as well. These findings are limited to the voiced stops /b,d,g/ and several other phonetic parameters still await investigation. Foremost is voicing. In particular, it is of some interest to determine if this analysis will correctly identify place for /ptk/ as well as /bdg/. Clearly, the two

primary features, Tilt of burst and Mid-frequency peaks, are located mostly in the burst portion of the spectral displays. Since this portion of the spectra should be very similar for voiced and voiceless consonants, we may confidently predict that these features will successfully identify place in /ptk/. The secondary feature of Late onset of F1, on the other hand, will clearly need some further modification. Only research with both the voiced and voiceless consonants will determine whether the Late onset feature can be defined in such a way that it can act as a reliable secondary cue for identifying /g/ in various contexts.

The features as defined in this study were used to identify place in what has been called initial stops. Initial in the context of this experiment means syllable initial since in fact all CV's examined here were originally extracted from the carrier sentence "Teddy said CV." No claims or hypotheses are being proposed on the basis of this experiment for presence of these features in running spectra of syllable final stops, or in running spectra of segments from other manner classes such as nasals or fricatives. For example, the Tilt of burst feature, as discussed, contains as part of its definition the location of a burst following a closure interval. Since this sequence of acoustic events is not found in nasals, the feature would not apply to nasals. Based on the acoustic theory of speech production, however, we would

expect that some aspects of the features would be observed in place features for stops in final position or for nasals. However, no specific proposals are offered here about the nature of these features.

The conclusions derived from the present experiment are quite different from the recent claims made by Stevens and Blumstein (1978; 1980). The invariant acoustic cues they propose in terms of onset spectra are linked to the general notion of distinctive features for place as proposed by Jakobson et al. (1952) and Chomsky and Halle (1968). Thus Stevens and Blumstein specifically claim that onset spectra can and should correctly specify place for final stops and nasals. However, their own research provides little evidence to support this claim. For final stops in the Blumstein and Stevens' template study (1979), the average correct identification of place at closure was 53%, and identification of the final burst (which is not typically present in running fluent speech) was 76%. In the preliminary study of [n] versus [m], average place identification was 76%. However, [n]'s were accepted by both the labial and alveolar template 67% of the time, so that the unique identification of [n]'s was only 9%. Similarly, unique identification of [m]'s was 71%. Therefore, the combined results of uniquely identifying [n] versus [m] was below 50% chance level for two choices. These results are quite poor and they cannot be used to support a strong claim that static onset spectra templates

can reliably capture the invariant properties for the distinctive feature of place in all environments or across several manner classes. Furthermore, we have argued that static onset spectra cannot adequately capture the acoustic information for place even in syllable initial stops because the temporal dimension has been eliminated. Thus, although the features in this experiment may be ultimately limited to identifying place in syllable initial, voiced and voiceless stops, it is fully expected that examination of auditory spectral representations of speech signals where the time dimension is properly preserved will also be successful in determining the acoustic correlates of other classes of speech sounds as well.

IV. EXPERIMENT 3. IDENTIFICATION OF STOPS AND VOWELS IN TRUNCATED NATURAL CV SYLLABLES

A. Introduction

A major conclusion of both the preceding experiment and Stevens and Blumstein's research (Stevens and Blumstein, 1978; Blumstein and Stevens, 1979) was that invariant acoustic information for specifying place can be found in the initial portion of a CV waveform. However, the investigation of spectral cues in Experiment 2 differs from Stevens and Blumstein's in two important respects. First, both investigations examined different durations of the stop consonant waveforms. The spectrum examined by Stevens and Blumstein (1978; Blumstein and Stevens, 1979) encompassed the first 20 ms of waveform, whereas the spectra examined in Experiment 2 encompassed the first 40 ms of waveform. Second, both investigations differed in the nature of the cues used to identify place. Stevens and Blumstein developed spectral templates which were visually matched to a single, static spectrum of a waveform, whereas Experiment 2 defined visual features which were based on dynamically changing spectral information observed in running spectra. Two perception experiments were conducted to examine these differences perceptually. The present experiment addresses the first issue, namely the duration of the stop waveform necessary to identify place of articulation accurately. The role of static versus

dynamically changing spectral information for specifying place will be examined in the next experiment, Experiment 4.

The primary goal of the first part of this experiment, Experiment 3A, was to determine the duration of the initial portion of a stop consonant syllable needed to identify place of articulation correctly. In particular, the study focused on whether there is enough acoustic information to specify place in the first 20 ms of a stop waveform as Stevens and Blumstein have suggested, or in the first 40 ms of the waveform as Experiment 2 has suggested. The second part of this experiment, Experiment 3B, examined the identification of both vowels and consonants in the same initial waveform segments. The goal of Experiment 3B was to investigate how consonant and vowel information is encoded and processed in syllable-sized acoustic units.

Aspects of the above issues have been addressed in numerous previous studies reported in the literature. One group of studies compared naive listeners' ability to identify place of articulation in the aperiodic (i.e., burst plus aspiration) portions of a CV waveform versus the voiced formant transition portions. Winitz, Scheib and Reeds (1972), LaRiviere, Winitz and Herriman (1975) and Ohde and Sharf (1977) all presented the aperiodic portions of /p, t, k/ paired with three vowels to subjects for identification of place. LaRiviere et al. and Ohde and Sharf reported very high levels of identification with

performance better than 90% in most cases. However, performance in the Winitz et al. (1972) study differed unexplicably from these results with percent correct identification in the 50% to 75% range. Nonetheless, both groups of investigators concluded that the aperiodic portion carried a great deal of acoustic information for identifying place. Although these studies encourage us to believe that invariant place information may reside in the initial portions of voiceless stop consonants, the durations of the test waveforms used in these studies varied unsystematically from 25 to 99 ms. Ohde and Sharf (1977) also studied the aperiodic portions of the voiced stops /b, d, g/ ranging in duration from 10 ms to 87 ms. However, they did not report the results separately for voiced and voiceless stops.

Tekieli and Cullinan (1979) conducted a study to determine the minimum duration necessary to correctly identify consonants and vowels from the initial portions of CV syllables. Their corpus recorded from one talker included /b,d,g,p,t,k,dʒ,tʃ/ each paired with eight vowels. They truncated each stimulus using an electronic gate at 10 ms intervals over the range of 10 to 150 ms. Trained phonetic students identified both the consonants and vowels. Detailed data analysis did not usually factor out place identification separately from voicing. However, from examination of their Fig. 2 it appears that place was accurately identified better than 95% correct for 30 ms durations averaged over all 6 stops.

Another group of studies employed tape splicing techniques to determine whether the aperiodic or voiced formant transitions carried the most salient place information for stop consonants (Schatz, 1954; Fischer-Jørgensen, 1972; Cole and Scott, 1974a; Dorman et al., 1977; Just, Suslick, Michaels and Shockey 1978). In this research, the aperiodic portion of a CV syllable was spliced onto a voiced segment, which was sometimes the same vowel and sometimes a different vowel. In general, the results of these studies, although not in complete agreement with each other, showed that the aperiodic portion was the predominant place cue only for certain vowel contexts. Unfortunately, it is not easy to interpret these results in terms of the questions set forth in this experiment. The tape splicing technique used in these experiments produced unnatural stimuli in which the continuity of spectral peaks for many stimuli was interrupted causing spectrally conflicting cues. It is interesting to note here that Dorman et al. (1977) concluded the bursts were effective cues to place when the spectral continuity of the burst and following vowel was preserved. Nevertheless, it is not easy to generalize the role of the aperiodic segment in providing reliable place information for the perception of natural stops from the results of these studies.

Experiment 3 was nearly completed when Blumstein and Stevens (1980) reported a series of experiments which

addressed many of the same questions examined in this experiment. In particular, they wanted to know if the initial portions of a stop CV syllable contained reliable information for consonantal place and vowel identification. Unfortunately, the stimuli and procedures employed in their study are subject to several methodological criticisms. First, Blumstein and Stevens conducted these experiments with synthetic stop consonant stimuli, /b, d, g/ and the vowels /i, a, u/. Although the full length synthetic stimuli were shown to be fairly good exemplars of /b, d, g/ in forced choice identification tests, they were only synthetic approximations to natural speech. In particular, the bursts in these stimuli were acoustically impoverished having only one formant of excited energy. Thus it is difficult to generalize their results to the identification of naturally produced CV's.

Second, Blumstein and Stevens (1980) indicated in their results that the shortest stimuli were 10 ms long. However, the total duration of the stimuli was considerably longer. All their results were presented in terms of the duration of only the voiced portion of the stimuli. Voicing durations varied from 10 to 46 ms plus an additional decay time of 20 ms during which the synthesizer resonances died out (Blumstein and Stevens, 1980, p. 650). The durations of the aperiodic waveforms ranged from 10 ms for /ba/ and /bu/ to 25 ms for /gi/. Aperiodic waveforms were not presented in isolation, so all stimuli contained a voiced waveform

portion. Thus, the shortest, so called 10 ms stimuli, varied in duration from 20 ms for /ba/ to 35 ms for /gi/, not including the decay time. Therefore, Blumstein and Stevens' (1980) conclusion "that information with regard to place of articulation for a voiced stop consonant resides in the initial 10-20 ms of a consonant-vowel syllable" is, in fact, quite misleading and is actually at variance with the true durations of the stimuli used in their experiments. In fact, of the 9 CV's studied, two of them, /bu/ and /gi/, were very poorly identified below the 46 ms durations with performance about 60% correct. Overall it is not easy to specify the minimum duration necessary to identify place from the results of the Blumstein and Steven's study.

Thus, an experiment to determine precisely how much place information actually resides in the early portions of natural stop consonant syllables is still needed. The present experiment examined this question with naturally spoken consonant-vowel syllables from two male talkers in five different vowel contexts. By computer editing techniques, the aperiodic and following waveform segments were carefully edited and measured. The experiment consisted of two main parts. In Experiment 3A subjects identified the truncated CV's at various durations for consonant place alone. In Experiment 3B another group of subjects identified <u>both</u> the consonants and the vowels in the truncated CV's.

B. Stimuli: Identification of consonants in full syllables

Before the set of stop consonant-vowel syllables was edited for Experiment 3, we checked that listeners could correctly identify the consonants in the full syllables. A subset of 30 CV's from Experiment 2 was used in the present experiment. For both male talkers, one token each of /b, d, g/ paired with /i, e, a, o, u/ was randomly selected from the three tokens used in Experiment 2. A computer program was used to randomize and output the full syllables through a 12 bit D/A converter for recording on audio tape. The tape consisted of 10 blocks of the 30 CV's for a total of 300 trials. There were 3 seconds between stimuli and 10 seconds between blocks. Six naive subjects listened to the tape over headphones in a quiet room and were paid for their services. Subjects were given instructions to write down the letter which corresponded to the consonant they heard at the beginning of each syllable. The response set, therefore, was the open set of all English consonants. Results showed that subjects identified the stop consonants in the full syllables 99.8% correct. No consonant responses other than b, d or g were used. These findings demonstrate that all 30 original CV syllables can be considered good exemplars of the stop consonant the talkers had intended to produce.

C. Stimuli: Waveform editing

Each of the 30 original CV's was then edited digitally to retain only the initial portions of the waveforms. Five different cuts were made at zero crossings to produce five truncated stops from each CV. For /d/ and /g/, the first cut was made just before the first voicing pulse. This aperiodic portion of the waveform, containing the stop release burst and aspiration, will be referred to as the burst. The second cut included the burst and the first pitch pulse. For /b/, it was not always possible to obtain a burst-only waveform portion because voicing was occasionally continuous from the carrier phrase, "Teddy said" into the voiced stop syllable. Thus the first waveform cut for /b/ included the burst and the first pitch pulse. The next cut included the burst and two pitch pulses. The remaining cuts for all stops were made so that the waveform segments included the burst plus 3, 5, or 7 pitch pulses. By this editing procedure, the durations of the shorter truncated stops for /b/ and /d/ were approximately the same length, but /g/ durations were slightly longer. The total number of test stimuli produced by this editing procedure was 150, with durations ranging from 6 to 111 ms in length.

After the data were collected in this experiment, the pattern of results suggested that one stimulus should be reexamined to see if a waveform editing error had occurred.

This stimulus was the burst plus one pitch pulse /da/ from speaker RP. Unfortunately examination on the CRT revealed that the last digitized point missed the zero crossing by about 30%. Although the resulting click could not be easily heard in the 19 ms stimulus, it was perceptible and appeared to interfer with correctly identifying it as alveolar.

D. Procedure

The present experiment involved the collection of two sets of identification data from the truncated CV's, one set for the consonants and the other for the vowels. Table 10 provides a detailed summary of the procedures used across testing days.

Insert Table 10 about here

Each procedure began with a brief familiarization task followed by the identification test. Only one set of audio tapes was used in both identification tasks of Experiment 3, although the familiarization tapes were arranged differently. All tapes were produced by a computer program which selected the digital waveforms on disk, randomized them and then output the stimuli through the D/A converter. The identification tapes consisted of six blocks of all 150 truncated stops. Stops were randomized within blocks, with three seconds between stimuli and seven seconds after each

Table 10. Procedures for consonant and vowel identification tasks for Experiment 3.

Consonant Identification

Day 1

	Tape	Task	Stimuli
1.	CON I	Cued familiarization	60 truncated, ordered
2.	CON II	Feedback familiarization	25 truncated, randomized
3.	ID	Forced choice ID (b,d,g,p,t,k)	450 truncated 3 blocks of 150 randomized

Day 2

1.	ID	Forced choice ID (b,d,g,p,t,k)	450 truncated 3 blocks of 150 randomized

Vowel Identification

Day 1

	Tape	Task	Stimuli
1.	VOW I	Vowel transcription training with feedback	30 original syllables, randomized
2.	VOW II	Cued familiarization	60 truncated, ordered
3.	VOW III	Feedback familiarization	25 truncated, randomized
4.	ID	Forced choice ID (EE,AY,AH,OA,UU)	450 truncated 3 blocks of 150 randomized

Day 2

1.	ID	Forced choice ID (EE,AY,AH,OA,UU)	450 truncated 3 blocks of 150 randomized

block of 50 stimuli. Two tapes were made for the familiarization task preceding consonant identification. The first tape (CON I) contained a subset of 60 of the 150 truncated stops, half from each talker. Five waveform portions edited from 12 original CV's were ordered from longest to shortest in duration with three seconds between items. The second tape (CON II) contained 25 additional truncated stops, several selected from each talker. The stimuli were randomized and presented at three second intervals.

Three different tapes were prepared for the familiarization task preceding vowel identification. The first tape (VOW I) contained each of the 30 original full syllables randomized once at three second intervals. The second tape (VOW II) consisted of a subset of 60 truncated stops, half from each talker. The waveform portions were ordered for presentation from longest to shortest for three cuts each from 20 original CV's. The last tape (VOW III) contained 25 additional stimuli, randomized and presented at three second intervals.

Each identification test was given on two days. Day 1 included the short familiarization tasks plus the forced choice identification test for the first three blocks of test trials. Responses were always recorded by hand on prepared answer forms. For consonant identification, the responses were (b, d, g, p, t, k). Although all stimuli used in Experiment 3 were edited from voiced stop

consonants, earlier pilot work indicated that naive subjects were more comfortable identifying the shortest waveforms with the voiceless stop responses (p, t, k). No special training was necessary to acquaint naive listeners with the stop consonant response set.

For vowel identification, the response set for /i, e, a, o, u/ was (EE, AY, AH, OA, UU). In order for the subjects to become familiar with these responses, a short vowel transcription task with feedback was conducted using the full syllables on tape VOW I. An index card with the correct vowel responses was in front of subjects throughout the experiment.

Two brief familiarization tasks preceded each identification test to acquaint the naive subjects with the truncated stimuli. In the cued familiarization task, subjects saw the correct response on their answer sheet before listening to the stimulus. In the feedback task, subjects heard a stimulus, wrote down a response, and then moved a slider to uncover the correct response. The total number of stimuli used in the familiarization tasks was 85 of the original 150 stimuli.

Subjects listened to the stimuli through TDH-39 earphones in a quiet testing room. Audio tapes were played back on an Ampex AG-500 taperecorder. A comfortable listening level for the brief stimuli was selected and a single repeated stimulus recorded on each tape was used to calibrate the listening level for all tapes. The level of

presentation of this repeated item was approximately 85 dB. Separate written instructions were given for each task. Subjects were contacted through a laboratory paid subject pool. Subjects were paid $6 for two days of testing in Experiment 3A and $13 for four days of testing in Experiment 3B. Subjects were phonetically naive and had no known hearing or speech disorder at the time of testing as assessed by a pretest questionaire.

E. Results: Consonant-only identification, Experiment 3A

Ten subjects participated in the consonant-only identification task. One subject skipped so many responses on the first day that she was asked not to return for the second day. Thus, the results for the consonant-only experiment were analyzed from nine subjects, providing 54 data points for each truncated stimulus.

Responses were scored as correct when place of articulation was correctly identified regardless of the voicing judgement. Collapsing over all responses and stimuli, subjects identified place of articulation correctly on 93.2% of all trials. This level of performance is surprising when we consider that over half of the stimuli were shorter than 45 ms. Relatively little effect of learning could be observed from Day 1 to Day 2, a change from 92.8% to 93.7% correct. Identification of the stimuli from the two talkers was the same at 93% correct overall.

Insert Figure 9 about here

Results for Experiment 3A are summarized in Fig. 9. Identification functions averaged across vowels are plotted separately for each stop and each talker according to the number of pitch pulses in the stimuli. The functions for each talker are clearly very similar and constitute an internal replication of the basic results.

Identification performance for /b/ as seen in Fig. 9 can be characterized as starting with an average of 90% correct identification for the burst plus one pitch pulse stimulus and rising to close to 100% correct with the next pitch pulse. The identification functions for /d/ are similar to /b/, but rise somewhat more gradually to 100% correct. The identification performance for /g/ differs from /b/ and /d/. For the burst-only segment, identification is not very accurate with performance at about 70% correct. Furthermore, /g/ identification functions never reach the 100% correct level for the longest stimuli, as do the /b/ and /d/ functions.

Insert Figures 10, 11 and 12 about here

To examine the relations between the durations of the truncated stops and the correct identification of place, the results are plotted separately for all 30 CV's in

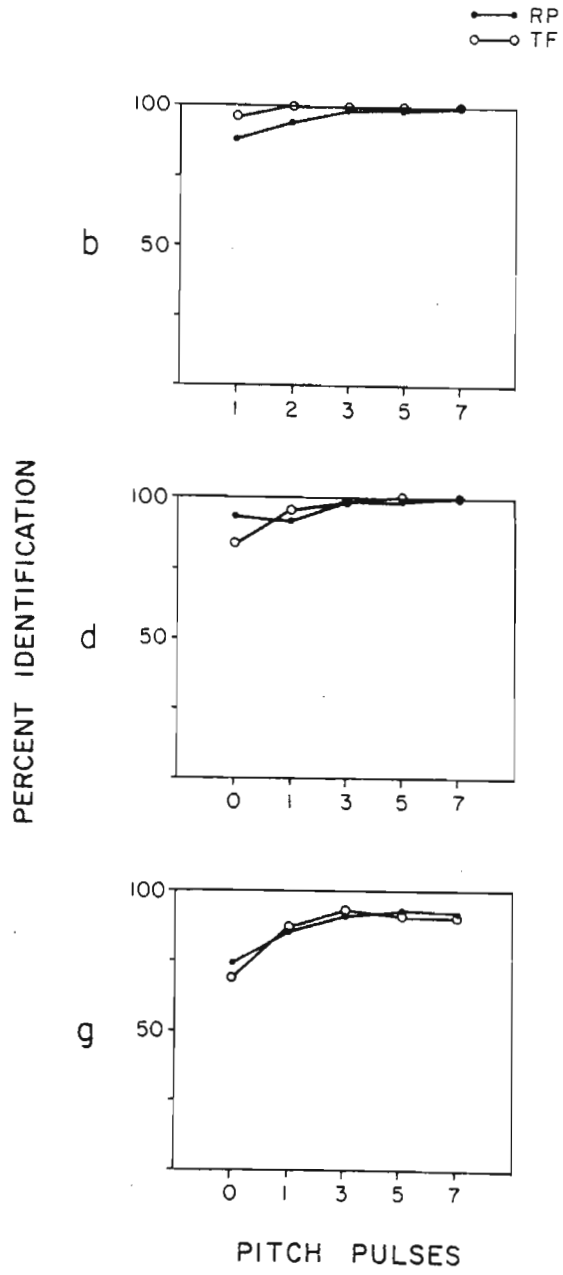


Figure 9. Percent correct consonant identification for stimuli containing increasing number of pitch pulses produced by two talkers, RP and TF. Each panel presents the data by consonant averaged over five vowels.

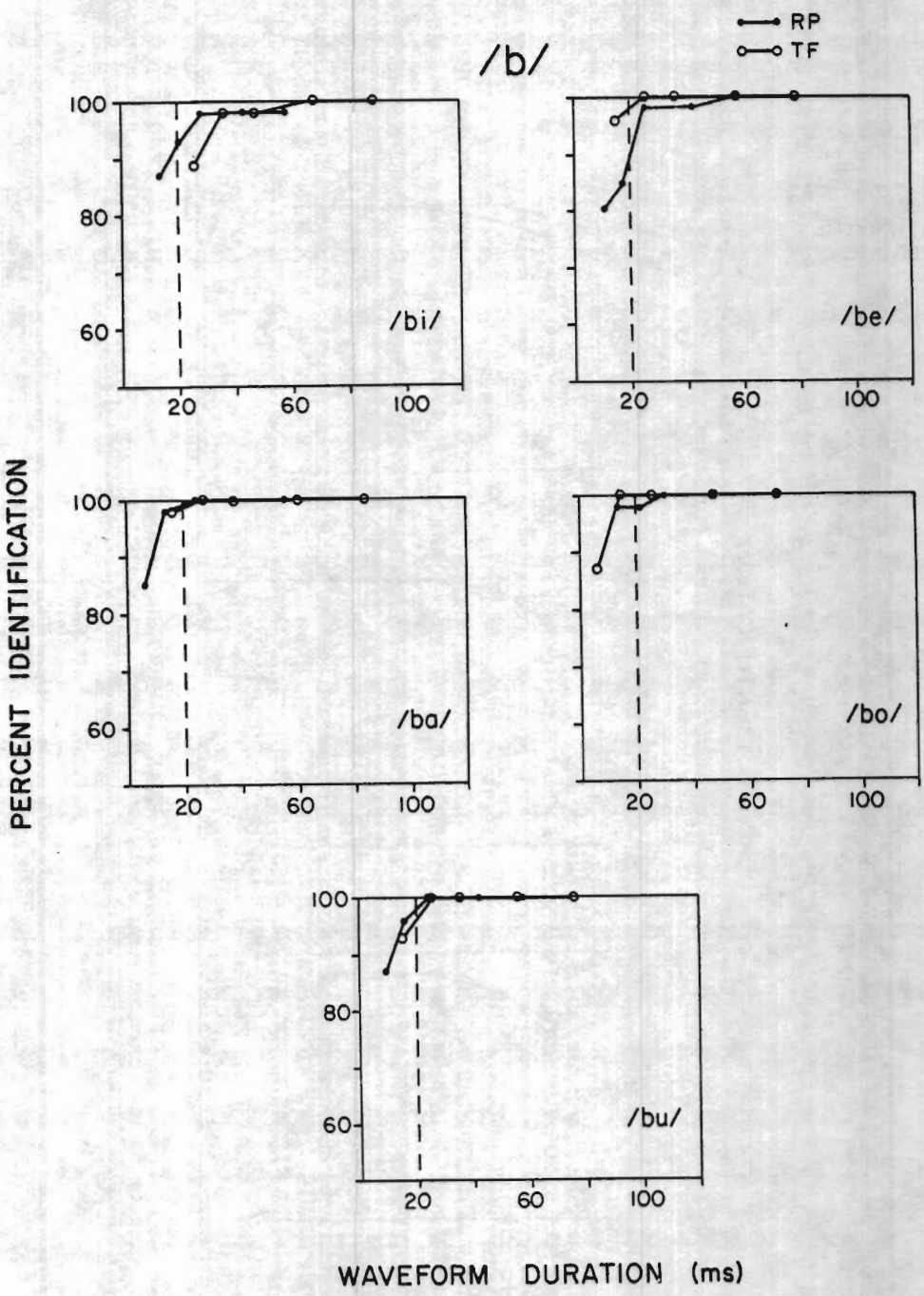


Figure 10. Percent correct consonant identification functions for all bilabial syllables produced by two talkers, RP and TF, are plotted by stimulus duration.

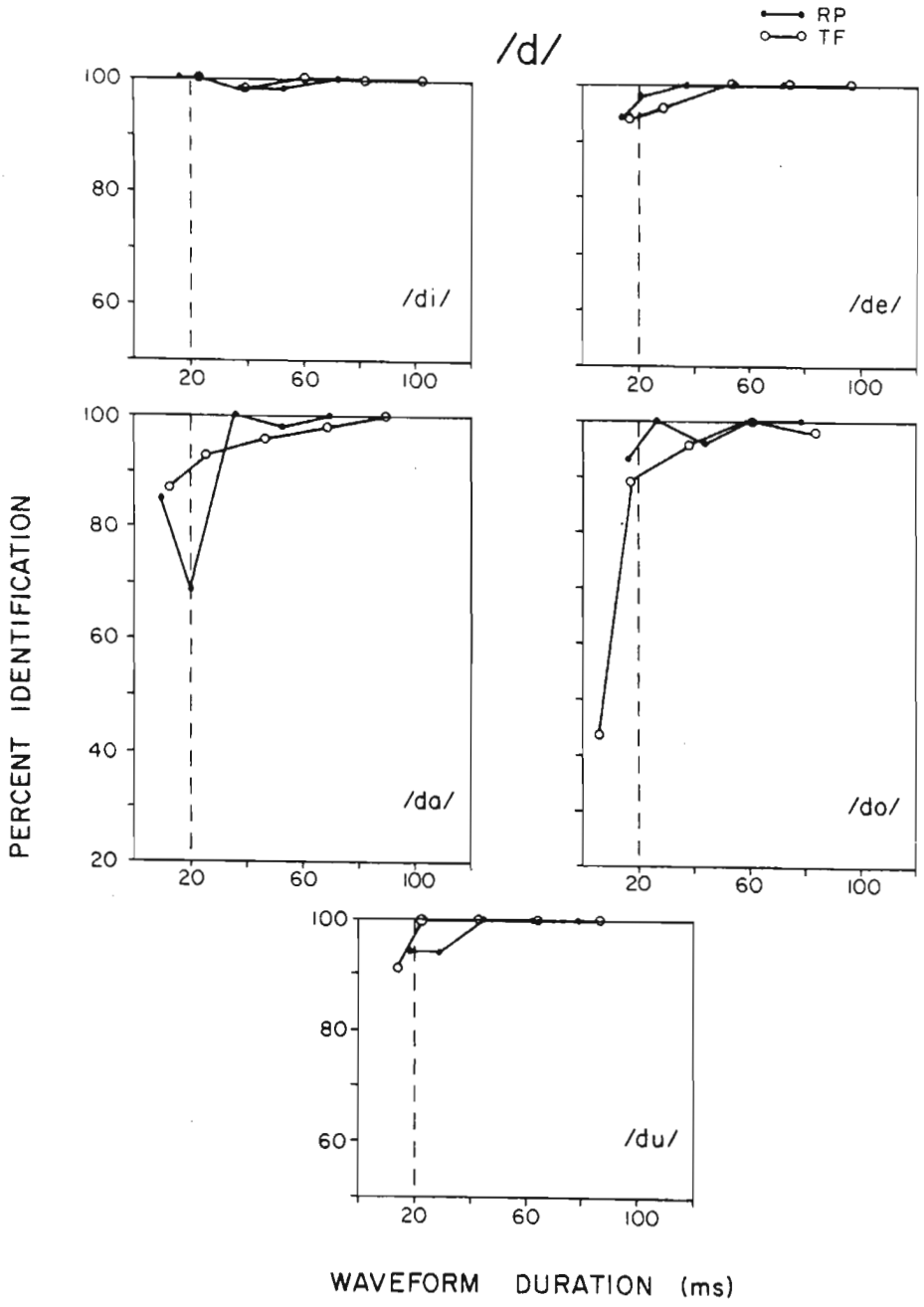


Figure 11. Percent correct consonant identification functions for all alveolar syllables produced by two talkers, RP and TF, are plotted by stimulus duration.

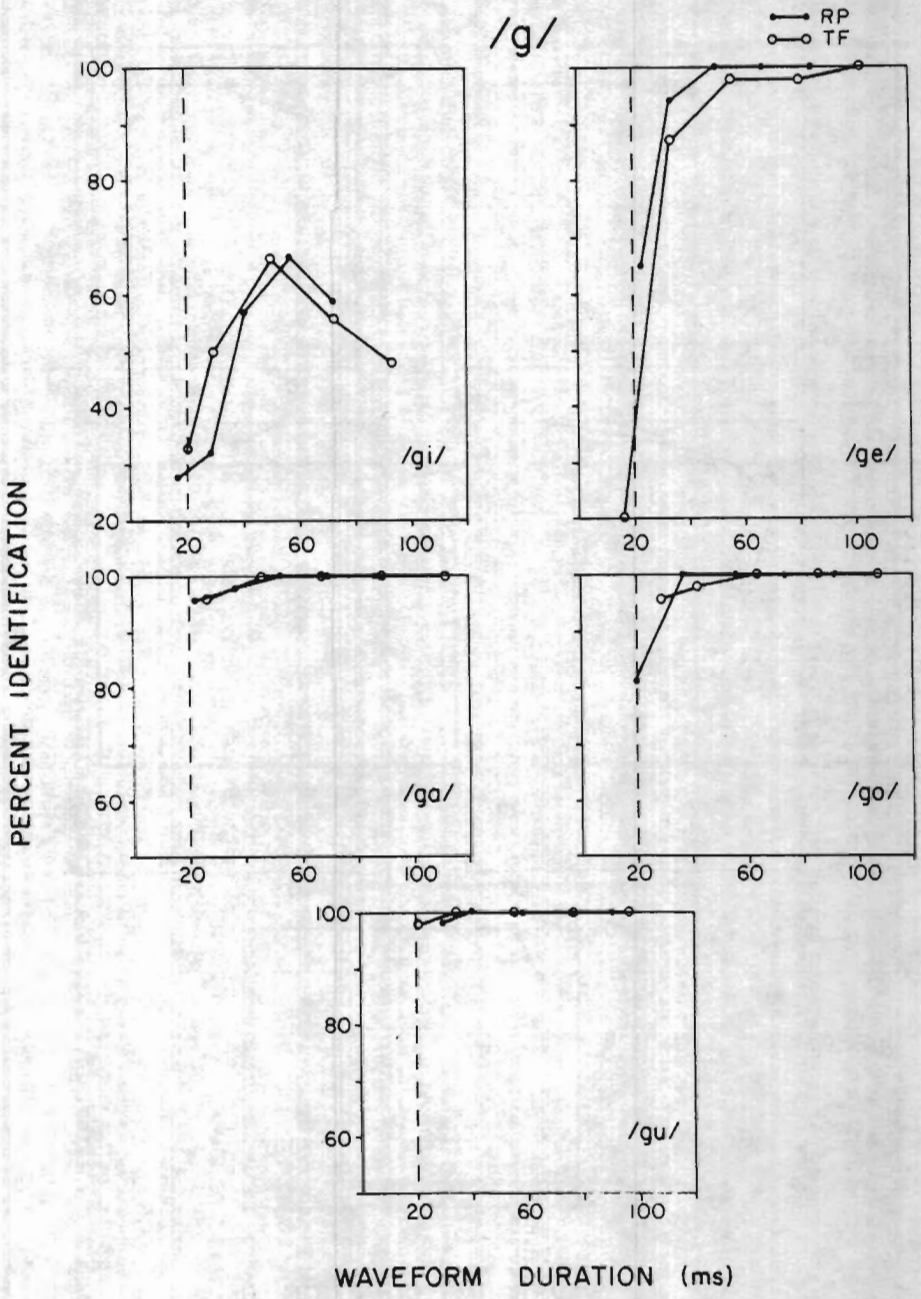


Figure 12. Percent correct consonant identification functions for all velar syllables produced by two talkers, RP and TF, are plotted by stimulus duration.

Figures 10, 11 and 12. Identification performance for all vowel contexts of /b/, shown in Fig. 10, was similar to the average functions plotted in Fig. 9. Individual functions for /d/ on Fig. 11 were also quite similar to the average /d/ functions with two exceptions. First, in Fig. 11 the /do/ burst was identified correctly only 44% for speaker TF. The short 6 ms duration alone cannot explain the poor /do/ identification since there were two /b/'s whose duration was also 6 ms but these were identified 86% correctly. The second exception was the /da/ stimulus from talker RP which had the only non-monotonic identification function (see the V-shaped function on Fig. 11 near 20 ms). Apparently, the waveform editing error described above reduced the correct identification of the alveolar stop.

The results for /g/ demonstrated substantially more vowel context dependency effects than /b/ or /d/ as can be seen in Fig. 12. The most unusual identification functions in Experiment 3A were obtained for the /gi/ stimulus. Identification was poor at all waveform durations and was less than 50% for the longest (93 ms) stimulus. The identification of /ge/ was also poor for the burst-only segments, but increased with three pitch pulses to better than 95% correct. The identification functions for the back vowels with /g/ were quite similar to those of /b/ and /d/. Apparently, /g/ before front vowels, with a more palatal place of articulation, was more difficult for subjects to identify than /g/ before back vowels, with a velar place of articulation.

To summarize, naive subjects were able to identify correct place of articulation in initial portions of natural stop CV syllables. The shortest waveform (burst-only for /d, g/ or burst plus one pitch pulse for /b/) was identified with greater than 80% accuracy for 26 out of the 30 CV's examined. The single exception was /gi/ which achieved less than 60% correct identification even with the longest duration stimuli.

Results from Experiment 3A can be used to evaluate the underlying principles of the Blumstein and Stevens (1979) template analysis, as well as the proposed running spectral features from Experiment 2. Blumstein and Stevens' templates were designed to fit spectral sections of the first 25.6 ms of a stop waveform, which after windowing, had an effective duration of 20 ms. Results from this experiment therefore can be used to evaluate whether or not sufficient information for identifying place of articulation lies within the first 20 ms of a stop consonant vowel syllable. A dashed vertical line has been drawn in Figures 10, 11 and 12 at 20 ms. The intersection of an identification function with the dashed line indicates the predicted response accuracy for identification of place from the first 20 ms of a test stimulus. The 20 ms identification values were estimated separately for all 30 CV's (smoothing over the bad /da/ stimulus for RP). These values were then averaged across vowels and talkers for each consonant and appear in the first column of Table 11.

Insert Table 11 about here

As shown in the table, the first 20 ms of a stop waveform contains sufficient place information for /b/ (96% correct) and /d/ (94% correct), but not for /g/ (73% correct). That is, a template matching analysis of the first 20 ms of spectrum might be expected to succeed for /b/ and /d/, but not for /g/. Notice, however, that the source of errors for /g/ was not uniform but occurred mostly for stimuli /gi/ and /ge/. In this study, the syllables /gi/ and /ge/ are representative of the front vowels in English. Therefore, there is little principled reason to treat them as simple exceptions to the Blumstein and Stevens invariant onset spectrum hypothesis.

The assumption underlying the feature definitions used in the running spectra experiment, Experiment 2, was that dynamic changes in the distribution of spectral energy over time can be used to specify differences in place of articulation. The most important feature that distinguished /b/ from /d/ in this three feature system was the tilt of the spectrum at burst onset. The identification of /b/ and /d/ also require the absence of extended Mid-frequency peaks. Reinterpreting this, the feature system requires that either no mid-frequency peaks exist in the early part of the waveform, or if they occur, as they sometimes do in /d/ bursts, that they dissipate rapidly. Because

Table 11. Percent correct identification of place in several different experiments averaged across vowels.

	20 ms identification Experiment 3A	50 ms identification Experiment 3A	Identification in running spectra, Experiment 2 (males only)
b	96	99	91
d	94	99	96
g	73	92	90

information about spectral tilt of the burst and each of the Mid-frequency peaks resides in the earliest portions of the stop waveform, this feature system implies that identification of /b/ and /d/ should be quite good for very short truncated stimuli. This hypothesis was confirmed by the high level of correct identification for /b/ and /d/ as shown in Table 11.

For /g/, the two proposed spectral features were Late onset of low frequency energy, and the presence of Mid-frequency peaks extending in time. Both feature definitions imply that more than 20 ms of a stop waveform is needed to identify /g/ correctly. The identification functions shown in Fig. 12 indicate that very high identification of /g/ was achieved at durations of 40 to 50 ms for all vowels except /i/. In order to quantify the identification performance of /g/ at longer waveform durations, 50 ms identification values were located on Figures 10, 11 and 12 in the same manner as the 20 ms identification values. Fifty millisecond values were chosen here because the duration of spectral information displayed in the eight frames of a running spectrum in Experiment 2 was close to 50 ms. The 50 ms identification values for each consonant were also averaged over vowels and talkers and are also presented in Table 11. These results indicate substantial improvement in identification of /g/ from the 20 ms point at 73% correct to the 50 ms point at 92% correct. Of course, only a small amount of improvement in

identification was observed for /b/ and /d/, since such a high level of performance was already achieved at 20 ms.

Since the 30 CV's used in Experiment 3 were a subset of those used in Experiment 2, a more direct comparison of the results from these two experiments is possible. The data for the identification of stops from the running spectral displays in Experiment 2 were averaged across the two male talkers and all vowels and are presented in the last column in Table 11. (The female talker's results were excluded from this average since her syllables were not used as test stimuli in Experiment 3.) Earlier in the discussion section of Experiment 2 we hypothesized that the running spectral features used to identify place should be salient spectral cues for the human auditory system in CV syllables. If correct judgements of place by listeners in Experiment 3A were to reach roughly the same level of performance as correct judgements by viewers of the running spectra in Experiment 2, for the same signal durations, the results could be interpreted as support for this hypothesis. A comparison of results in Table 11 indicates that the identification of place in both experiments is quite similar for /d/ and /g/. Both experiments produced high levels of identification within 3% of one another. The identification levels for /b/ in the two experiments were not as close, 91% compared to 99%. In the Discussion of Experiment 2 we have already made some suggestions for improving the definition of the Tilt of burst feature used

in identifying /b/. These suggestions included redefining the tilt for /d/ as rising prominently and a simple rule for accounting for differences in absolute frequency due to vocal tract size.

Of further interest to this comparison are the sources of errors in the two experiments. In both experiments most of the errors occurred on /g/. More specifically, the bulk of the identification errors occurred for the high front vowels of /gi/ and /ge/. Thus this comparison has shown that identification of place of articulation by visual inspection of dynamically changing spectral information is in fairly good agreement with listeners' identification of the same place features in the present experiment over approximately the same initial waveform durations.

To summarize, the results from naive listeners' identification of truncated CV syllables showed that /b/ and /d/ could be identified at 94% or better given only the first 20 ms of the waveform. From this we can conclude that either the Stevens and Blumstein (1978) onset spectrum approach, or the running spectrum approach of Experiment 2 could succeed in specifying invariant place information in CV syllables. For /g/, however, results from this experiment indicate that velar stops are poorly identified (73%) from the first 20 ms of waveform, and that considerably longer waveform durations are needed for accurate identification. From this result we conclude that the Stevens and Blumstein approach probably cannot specify

velar information in a single 20 ms, static onset spectrum. On the other hand, the running spectral features for identifying /g/ were specifically defined to characterize changing spectral information over the first 40 ms of a waveform. A detailed comparison of the identification and error results obtained in Experiments 2 and 3 showed that performance in both cases was in close agreement. Apparently, the underlying principles of the dynamically changing spectral features of Experiment 2 are more likely to capture reliable information for place of articulation than the fixed 20 ms static onset spectra of Stevens and Blumstein.

F. Results: Stop versus Vowel Identification, Experiment 3B.

The results of the consonant-only identification tests in Experiment 3A revealed that listeners were very accurate in identifying place of articulation in truncated CV syllables. During the course of this work it became of some interest to determine how well vowels could be correctly identified in these same stimuli. A comparison of the identification functions for consonants and vowels obtained from the same group of listeners could provide several insights into the way acoustic-phonetic information is encoded and processed in CV syllables. This issue seemed appropriate in this context because for many years Liberman and his colleagues (Liberman et al., 1967; Liberman, 1970;

Liberman, Mattingly and Turvey, 1972; Liberman and Studdert-Kennedy, 1978) have argued that the acoustic information for identifying consonant and vowel segments in a CV syllable is encoded simultaneously throughout the whole syllable. Liberman has referred to this in the literature extensively as the parallel transmission of acoustic cues (Liberman et al., 1967). The consonant and vowel identification functions of truncated syllables could provide some interesting new evidence to evaluate these claims. For example, consider the shape of the average consonant identification functions in Fig. 7. If consonant and vowel information is encoded simultaneously in CV syllables, we would predict that vowel identification functions obtained from the same stimuli would be roughly parallel to those seen in Fig. 7 for the consonants. On the other hand, if the vowel identification functions demonstrated much poorer performance or varied in uncorrelated ways with the data obtained for consonants, the claims for parallel transmission of acoustic cues would not be supported.

Blumstein and Stevens (1980) in their study of short synthetic stops included a vowel identification condition similar to the one examined in this experiment. Their basic argument was that the consonantal cues are derived from the onset spectra separately from the vowel cues located in the formant transitions (Blumstein and Stevens, 1980, p. 660). However, at the conclusion of their vowel study they made

several statements which partially agree with the earlier position of Liberman and his colleagues on parallel transmission. Specifically, Blumstein and Stevens hypothesized that a short stimulus is perceived holistically as a unitary syllable, containing both consonant and vowel features. The present experiment using natural speech and more CV tokens was designed to investigate this claim in greater detail.

An additional, closely related question, also examined in this experiment, was whether the perceptual processes for extracting the consonant and vowel information are independent or dependent on one another. Liberman (Liberman et al., 1967; Liberman et al., 1972) has argued that the perception of consonant and vowel information is achieved by parallel, dependent perceptual processes in the same way that consonant and vowel information is encoded in parallel throughout the entire formant trajectories into the vowel nucleus. Blumstein and Stevens (1980) expressed very much the same point of view. They suggested that the acoustic correlates of the consonant and vowel are processed in parallel, and not processed by independent, sequential operations. The comparison of the shapes of the consonant and vowel identification functions therefore should provide some insights here as well. If the perceptual processes are dependent, the consonant and vowel functions should not only be correlated, but they should also show similar levels of identification performance at any particular

waveform duration. That is, if the identification of the vowel depends in some way on consonant identification, then performance levels for vowel and consonant identification should be roughly similar for the same stimulus duration in order for the vowel processes to obtain some beneficial information from consonant identification. On the other hand, if the processes are independent, consonant and vowel functions should show substantially different performance levels for the same waveform segment.

A replication of Experiment 3A with the addition of the vowel identification task (see Procedure) was conducted to answer these questions. Four groups of six subjects each were contacted through the laboratory paid subject pool to participate in the four day experiment. Two groups were randomly assigned to take the consonant identification test first: the other two groups took the vowel identification test first. Subjects participated in one identification test on successive days and then returned the following week on the same two days to participate in the other test. Subjects were not told in advance that there would be two separate tasks.

After completion of Experiment 3A, it was decided that all subjects should be screened audiometrically for normal hearing before being required to identify such extremely brief stimuli. Therefore, before testing in Experiment 3B, all subjects were required to pass an audiometric test for the octave audiometric frequencies from 500 to 8000 Hz at a

sound pressure level of 20 dB (ANSI-1969) using a Grason-Stadler Model 1701 audiometer.

One subject did not pass the screening test, and numerous other subjects were unwilling to complete the four day testing procedure. At the end of testing, a complete set of data was obtained from seven subjects in the consonant-first group and nine subjects in the vowel-first group.

The consonant-first group may be considered as a replication of Experiment 3A, except for the addition of the audiometric screening procedure. Overall, correct consonant identification for this group was 90.5% compared to 93.2% in Experiment 3A. Subject variability in this task was considerable, ranging from 3% to 14% errors. However, the mean error rates for subjects in the two groups were not significantly different from one another by a t-test for independent samples, $\underline{t}(14) = 1.477$. Group identification functions for individual CV's in the replication Experiment 3B were virtually identical to those obtained earlier in Experiment 3A.

A comparison of the two subject groups in this experiment showed that the vowel-first group outperformed the consonant-first group in both identification tasks by an average of 7%. The vowel-first group was 5% better in identifying consonants and 9% better in identifying vowels. In debriefing after the experiment was completed, subjects who took the consonant-first condition complained of

difficulty in orienting to the vowel task. The vowel-first group voiced no complaints and found the consonant task quite easy (95% correct overall). The differences in the identification functions between the two groups can be characterized as a general upward shift of the vowel-first functions without changes in overall shape. Therefore, results for all sixteen subjects were combined for the final analysis.

Insert Figure 13 about here

Summary results for this experiment are plotted in Fig. 13. The results, shown separately for each talker, display consonant and vowel performance as a function of pitch pulses averaged over the 15 CV's. The results obtained from the two talkers show very similar identification functions and therefore constitute an additional replication of the previous results obtained in Experiment 3A. Figure 13 clearly shows that vowel identification was consistently poorer than consonant identification. The overall difference was 24%. Vowels were barely identified above chance for the shortest truncated stimuli, 39% correct, whereas consonants were identified better than 80% correct. A considerable discrepancy can be observed in performance between vowel and consonant identification functions even at the longest stimuli durations. For the longest stimuli (seven pitch pulses),

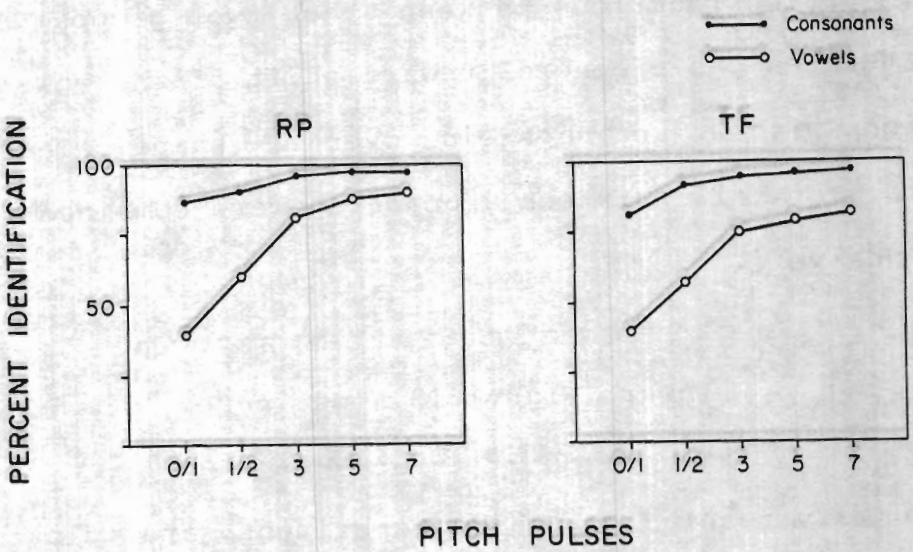


Figure 13. Percent correct identification of consonants and vowels displayed by number of pitch pulses. Data on each panel is for talker RP or TF averaged over three consonants and five vowels. The number of pitch pulses in the first two truncated stimuli for /b/ was 1 and 2, whereas for /d/ and /g/ it was 0 and 1.

consonants were identified better than 95% correct for 53% of the syllables, whereas vowels were identified 95% correct for only 10% of the syllables. Thus, from these overall results we can conclude that the identification of vowels in truncated CV's was substantially more difficult than consonants. This conclusion is also consistent with the subjects' informal comments concerning the difficulty of identifying the vowels mentioned previously.

Although Fig. 13 shows the overall trend of the results in this experiment, many of the individual CV's were not similar to the averaged data. Figures 14 to 18 display the individual consonant and vowel identification functions for each CV for each talker as a function of the number of pitch pulses.

Insert Figures 14 to 18 about here

With these figures, the issue of whether consonant and vowel information is simultaneously encoded throughout the truncated CV's can be examined. Consider first the shortest truncated stops for /d/ and /g/ which contained only aperiodic waveform segments. Except for the /i/ syllables, vowel identification was typically very poor and close to chance. In these same segments, however, consonant identification for several CV's (e.g., /go/ for RP and /da/ for TF) was better than 90% correct. Thus, there was an asymetry between consonant and vowel identification for the

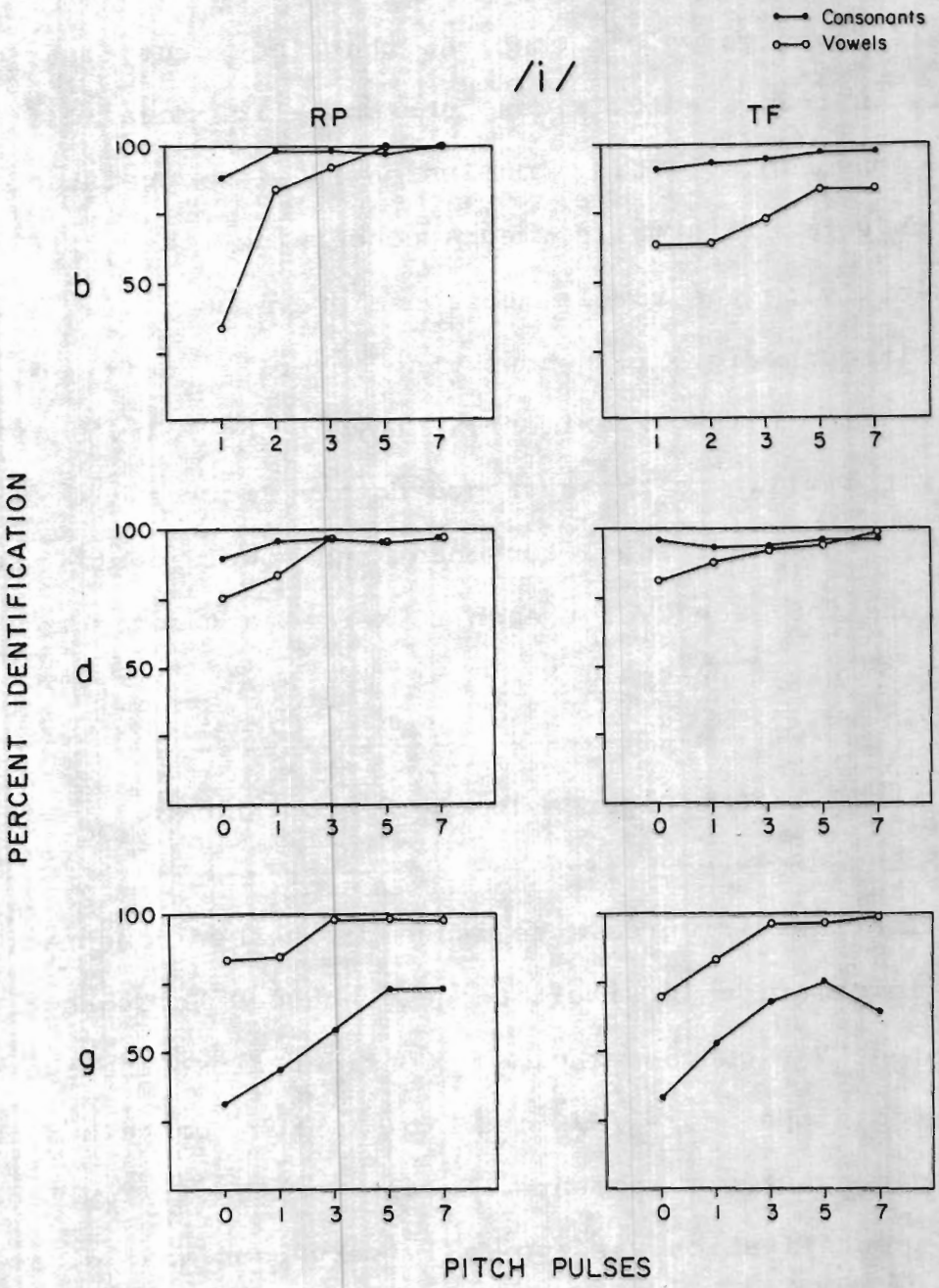


Figure 14. Percent correct identification of the consonants and vowels for all /i/ syllables plotted by number of pitch pulses separately for each talker, RP or TF.

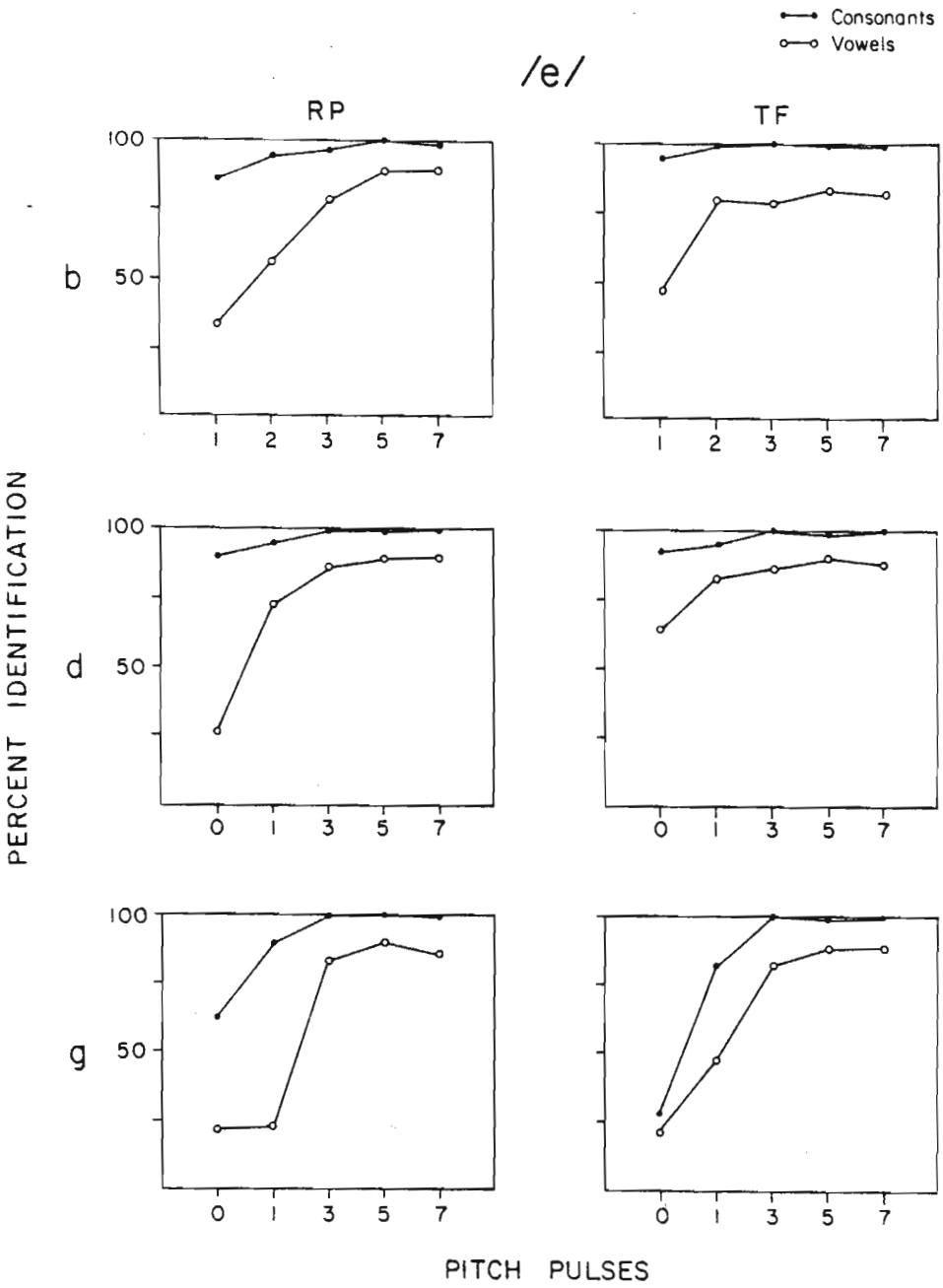


Figure 15. Percent correct identification of the consonants and vowels for all /e/ syllables plotted by number of pitch pulses separately for each talker, RP or TF.

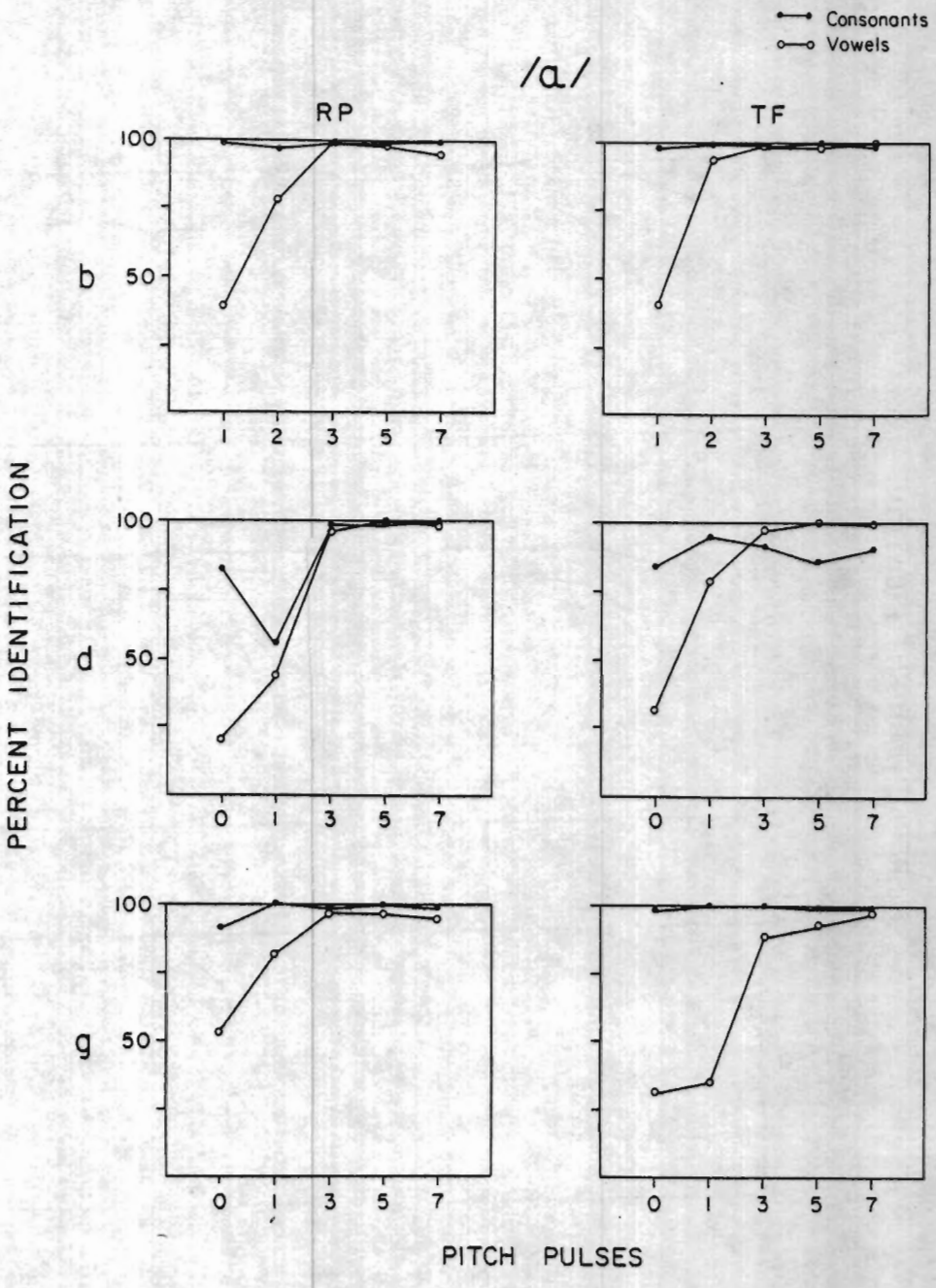


Figure 16. Percent correct identification of the consonants and vowels for all /a/ syllables plotted by number of pitch pulses separately for each talker, RP or TF.

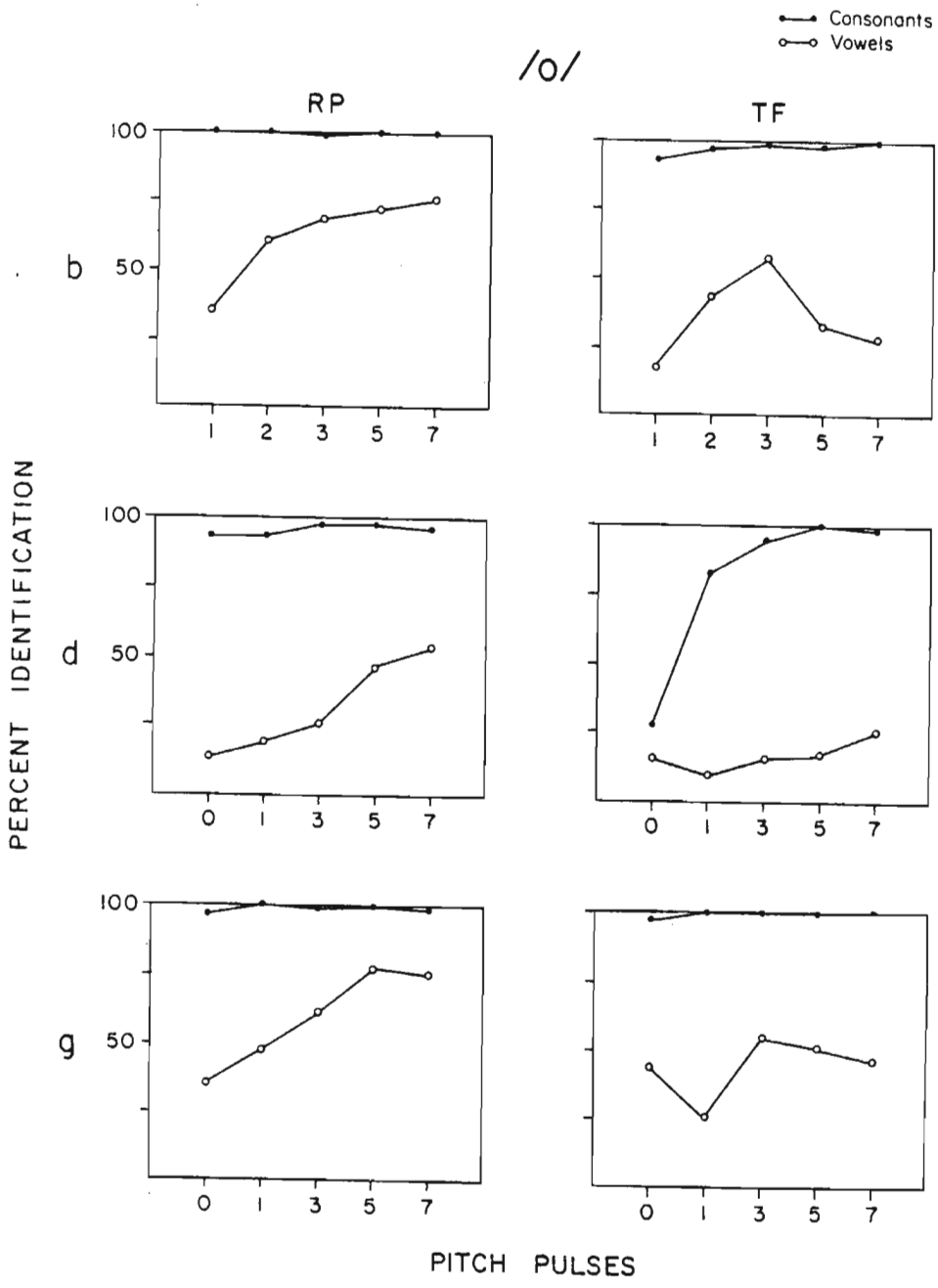


Figure 17. Percent correct identification of the consonants and vowels for all /o/ syllables plotted by number of pitch pulses separately for each talker, RP or TF.

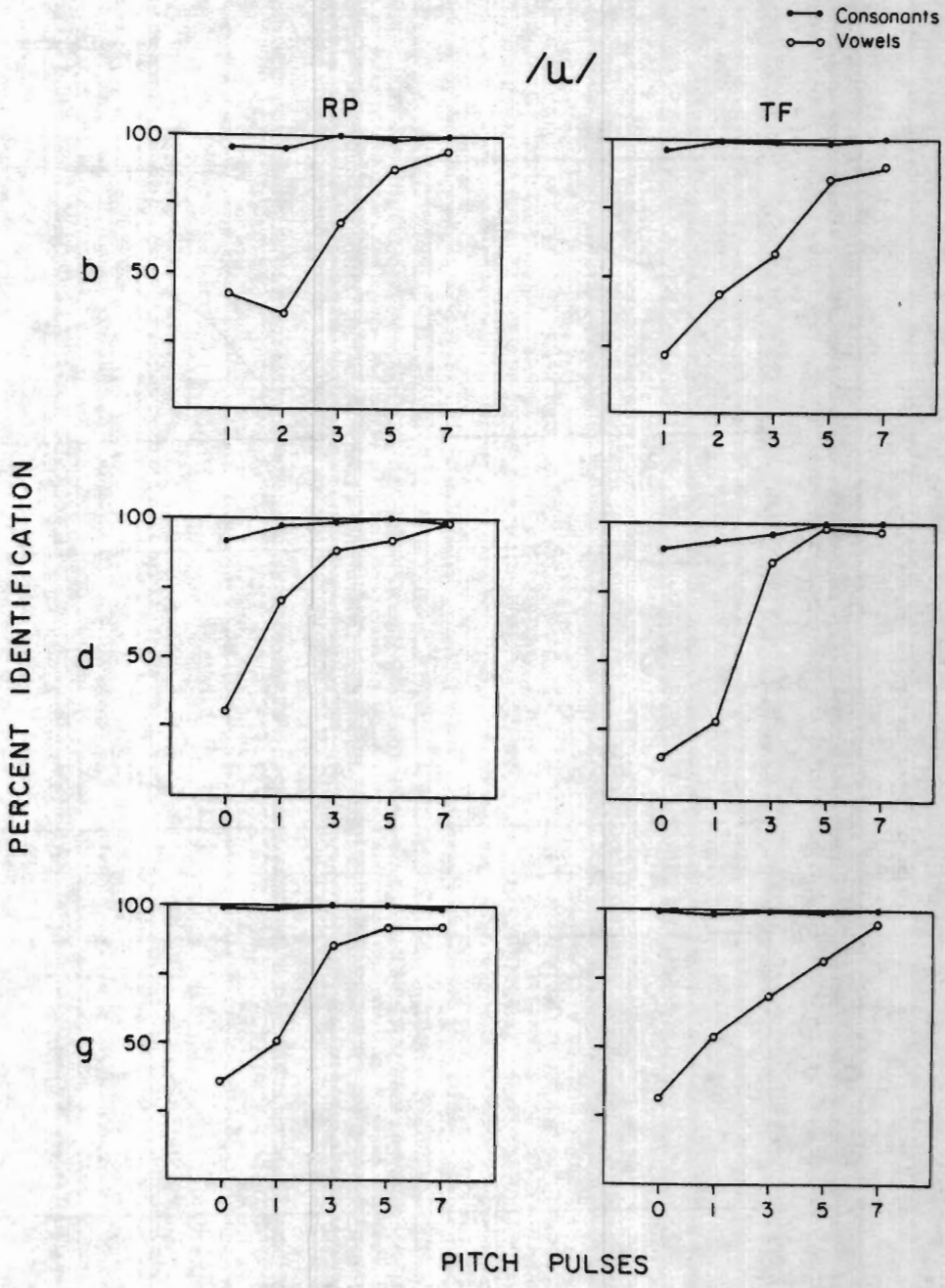


Figure 18. Percent correct identification of the consonants and vowels for all /u/ syllables plotted by number of pitch pulses separately for each talker, RP or TF.

short duration stimuli. Furthermore, not only was the vowel identification close to chance, but the vowel identification errors were random and not related to the acoustic similarity of the neighboring vowel responses. These results also agree with the earlier findings of Cullinan and Tekieli (1979) who reported only 35% correct vowel identification from the aperiodic segments of /p/, /t/, and /k/ paired with eight vowels. Thus, with the exception of the results for /i/, there is no evidence for parallel encoding of vowel information in the first aperiodic portion of a CV syllable.

At this point it should be noted that the parallel encoding hypothesis advocated by Liberman and his collaborators was derived from perceptual experiments with synthetic stop consonants having no release bursts. Based on these experiments it was assumed that the formant transitions encoded all the necessary and sufficient information for identifying place of articulation in stop consonants. Thus, although Liberman (Liberman et al., 1972) has generalized the parallel encoding hypothesis to mean that consonant information is encoded throughout the duration of natural CV syllables, perhaps the hypothesis only applies to the voiced segments of the waveform containing the formant transitions.

To test this hypothesis we determined how much improvement in consonant identification can be achieved by the addition of the voiced formant transitions to the

aperiodic segments. Only /gi/, /ge/ and the 6 ms /do/ for TF showed a substantial improvement in place identification in the voiced segments. Apparently, formant transitions do carry necessary information about place of articulation for /g/ with front vowels. But consider further the case of extremely brief bursts, such as those for TF's /do/, as well as those for all /b/ syllables whose shortest stimulus always contained one pitch pulse in this experiment. In these cases, it appears that the addition of one pitch pulse to the burst resulted in high consonant identification. However, in the case of low and back vowels, like our /do/ example, one pitch pulse is not a large proportion of the total formant transition length. Thus, if a burst is very short, the addition of a pitch pulse produces a longer signal which improves place identification, but the addition of more formant transition information beyond the first pitch pulse does not significantly improve place identification. This claim is supported by the observation that the average percent correct place identification over all stops, vowels and talkers for one-pitch-pulse stimuli was very high (91%). In short, it appears that in most naturally produced CV's, sufficient information for identifying place of articulation is encoded in waveform segments which precede the excursion of the formant transitions. These findings suggest that formant transitions do not in fact encode information necessary for place identification in many CV

syllables. This conclusion contrasts strongly with the position advocated by Liberman et al. (1967) that place of articulation information is encoded primarily in the formant transitions. At the same time, this conclusion is consistent with the results of Experiment 1 which found little evidence of context-invariant acoustic information in the formant transitions for reliably classifying place of articulation. Further implications of these findings will be developed in the discussion section below.

The emphasis of the Liberman position on parallel transmission, however, was that vowel information is also encoded in the formant transitions, since it was assumed that formant transitions carried the necessary and sufficient information for consonant identification. As mentioned previously, the present results indicate that virtually no vowel information was present in the aperiodic segments of the early portions of the stop consonant waveforms. By observing the change in the vowel identification functions during the voiced segments, the role of formant transition information in vowel identification can be determined. As shown in Fig. 13, most of the improvement in the vowel identification occurred during the first three pitch pulses which for most CV's encompasses the formant transitions. Therefore, vowel information does appear to be clearly encoded in the formant transitions of CV syllables. The individual vowel identification functions follow the average functions in

Fig. 13 except for /o/. The vowel /o/ was not accurately identified in segments containing less than five pitch pulses for either talker. However, it was the only strongly diphthongized vowel in the set /i,e,a,o,u/ examined in this experiment. In particular, for speaker TF, the onset of /o/ was low in frequency, rather close to /ʌ/. Thus, it is not surprising that naive listeners had considerable difficulty identifying /o/ in the truncated syllables. The response errors, however, demonstrated that naive listeners did extract vowel formant information since the vowel errors for the longer stimuli for both talkers were not random. Instead, error responses were acoustically similar to the vowel onsets of the diphthonized /o/. About three-fourths of the /o/ errors were /a/ responses; the remaining errors were /u/. Thus, results from the vowel identification task indicate that the acoustic information for vowels is encoded in the formant transitions of CV syllables. However, the correct phonemic identification of a diphthongized vowel like /o/ requires considerably longer voiced segments than only the information contained in the formant transitions.

To summarize, the results from Experiment 3B do not suppport the earlier claims of Liberman and his colleagues that consonant and vowel information is simultaneously encoded in parallel throughout a stop consonant-vowel syllable. Except for /g/ before front vowels, place information appears to be encoded in the aperiodic waveform

segment (or for very short bursts, the burst plus one pitch pulse.) The voiced formant transitions do appear to carry vowel information, but they carry substantial amounts of place information for only some of the syllables studied. The vowel information is generally sufficient to specify the location of the formants, but in the case of diphthongs like /o/, a considerably longer duration of the waveform is needed.

These results also contradict the earlier claim of Blumstein and Stevens (1980) that the truncated stops are perceived as unitary syllables. Acoustic information for specifying vowel identity was not available in the shortest segments used in this study. Furthermore, vowels were not accurately identified in one pitch pulse segments as reported by Blumstein and Stevens (1980, p. 658). Rather, for most vowels, accurate identification was only achieved for segments containing at least three pitch pulses. The superior vowel identification performance in the Blumstein and Stevens (1980) study may be due in part to task differences. Blumstein and Stevens employed phonetically trained listeners to identify vowels from a set of only three maximally-discriminable, synthetic vowels, /i,a,u/. This contrasts both with the present experiment and the earlier related experiment by Tekieli and Cullinan (1979). The present experiment employed naive listeners who identified vowels from a set of five naturally produced vowels, /i,e,a,o,u/, produced by two talkers. Tekieli and

Cullinan (1979) and Cullinan and Tekieli (1979) asked trained subjects to identify vowels in truncated waveform segments using eight English vowels. For isolated vowels, about 30 ms or 3 pitch pulses was needed for correct identification of the vowel according to their criteria (66% correct). For vowels preceded by /p,t,k/, vowels were identified in the longest aperiodic segments (average duration equal to 66 ms) only 39% correct. Apparently identification of vowels in brief waveforms using a larger set of vowels requires substantially longer waveform durations for comparable levels of performance.

The results described above can be examined further to determine whether the perceptual processes deployed in extracting the consonants and vowels from CV syllables are dependent or independent. The identification functions shown in Figures 14 to 18 strongly suggest that the acoustic cues for consonant and vowel identification are processed independently. Except for /i/, consonant identification exceeded 90% within one pitch pulse although vowel identification barely reached 44% correct. Clearly, accurate consonant identification could not be dependent on the poor vowel identification in these syllables. In addition, /gi/ provides a good example of the independence of vowel processing from consonant identification. The /i/ in /gi/ (and in /di/) was identified much more accurately (75%) for the aperiodic segments than the vowels in any other CV's, yet the consonant /g/ in /gi/ was the most

poorly identified consonant in the entire experiment. Thus, this result supports an account that emphasizes independent processing of consonant and vowel information in CV syllables.

The results of this experiment are therefore also in conflict with the dependent processing hypotheses of both Liberman (Liberman et al., 1967; Liberman, 1970) and Blumstein and Stevens (1980). In fact, they tend to support a processing strategy that Blumstein and Stevens (1980, p. 660) specifically rejected in their earlier work. Referring to the summary statement above, the results of Experiment 3B showed that consonant identification precedes vowel identification for most syllables, suggesting that processes for identifying consonants and vowels are largely sequential and independent operations. However, the results for the syllables /g/ with front vowels represent a partial exception to this hypothesis. The consonant /g/ was poorly identified in these short waveform segments and substantial durations of vowel information were needed for accurate place identification. Thus, the processing of consonant and vowel information in this one case is not sequential. On the other hand, the overall results suggest that the processes required for identifying /g/ are independent from the processes used to identify vowels. In the syllable /gi/ the vowel is accurately perceived, while in /ge/ the vowel is poorly perceived. A more definitive account of the nature of these processes awaits a further study of a

larger set of natural CV syllables than was used in this experiment. However, it is clear that the earlier claims made by Liberman et al. are not supported by the present findings using naturally produced CV syllables.

G. Discussion

The major outcome of this experiment indicates that sufficient acoustic information for identifying place of articulation in stop consonant syllables resides in the initial portions of the CV waveform. For /b/'s and /d/'s, place was correctly identified 94% for only the first 20 ms of the waveform. A longer duration, of 40 to 50 ms, was required for the correct identification of velar stops, particularly before front vowels. These results are consistent with the acoustic features proposed earlier in Experiment 2 for identifying place from visual representations of running spectral displays of the same waveforms. The critical features for /b/ and /d/ were Tilt of burst and absence of Mid-frequency peaks: these features are observable in the earliest portions of a stop waveform. The critical feature for /g/ was the presence of Mid-frequency peaks extended over time, with a secondary feature of Late onset of F1. The definition of the features for /g/ require somewhat longer portions of a stop waveform extended in time.

The finding that longer waveform segments were necessary to identify velars in this experiment is

incompatible with the Blumstein and Steven's (1979) onset spectra analysis which does not incorporate the temporal dimension. In this connection, it is interesting to note that Blumstein and Stevens were aware of this problem with velars although they made little effort to modify their feature definitions:

> a velar tends to be identified with fewer errors if the duration of the stimulus is longer than 10-20 ms, suggesting that a longer time is necessary to build up a representation of a "compact" onset spectrum in the auditory system. (1980, p. 660)

Although Blumstein and Stevens acknowledged that listeners may need extended duration to identify the compact feature, they proposed a static analysis procedure using onset spectra with no time dimension. Furthermore, they assumed that humans perceive the place features in stops by means of innate property-detectors specifically tuned to detect the overall gross shape of the spectra at onset (Blumstein and Stevens, 1979, p. 1014). However, this hypothesis leads to the following apparent contradiction in their reasoning. Although the compact feature must be specified over time in the auditory system, Stevens and Blumstein argue that it is perceived by property-detectors which are not sensitive to changes over time. Perhaps Blumstein and Stevens intended their conclusion to be interpreted in terms of some psychophysical property such as energy instead of time. It is not clear to us, however, what an interpretation other than time would mean since the experimental results of the present study clearly demonstrated the importance of duration for velar identification.

The second part of this study also revealed that sufficient information for correctly identifying consonants and vowels in CV syllables is not encoded in parallel over the same acoustic segments of the waveform. Consonants were identified accurately (about 90%) from the burst (aperiodic) or burst plus one pitch-pulse portion of the waveforms with the exception of /gi/. In these same segments, vowel identification was at chance in four of the five vowels studied. Vowels were identified accurately in waveform segments containing about three pitch-pulses, unless the vowels were diphthongs. The duration of the first three pitch-pulses was shown previously to encompass the duration of most of the measured formant transitions. Thus it appears that formant transitions do carry considerable information for identifying vowels in CV syllables. This conclusion concerning the role of formant transitions in vowel identification is virtually identical to the views expressed by Lindblom and Studdert-Kennedy (1967), Strange, Verbrugge, Shankweiler and Edman (1976), and Ohde and Sharf (1977). Formant transitions did not, however, provide much additional information concerning the identification of place of articulation in the stops since the waveform portions preceding the transitions already contained sufficient distinctive information to identify place of articulation.

The results of this study bear on a number of previous hypotheses concerning the role of the burst and formant

transitions as acoustic cues for specifying place of articulation in stop consonants. Essentially four positions have been proposed in the literature. First, the burst (aperiodic waveform) specifies invariant place information (Cole and Scott, 1974a). Second, formant transitions specify context-dependent place information (Liberman et al., 1967). Third, the burst and formant transitions both contain acoustic information for place of articulation in a "cue-trading" relation depending on the vowel context (Dorman et al., 1977). Fourth, place information is derived from both the burst and the formant transitions as a single, integrated cue (Fant, 1960; Stevens and Blumstein, 1978). The remainder of this discussion will review these hypotheses in light of our findings. Finally, a somewhat different proposal for the "direct" perception of consonants and vowels in initial stop syllables will be outlined to account for the present results.

The first three hypotheses assume that the burst is an acoustic cue which is independent and separable from the voiced formant transitions in CV syllables. This assumption is based primarily on visual inspection of oscillographic traces of stop-vowel waveforms and their spectral representation on sound spectrograms. However, this basic assumption may be artifactual and perhaps even false. Nonetheless, it has been the basis of several long-standing hypotheses concerning the acoustic cues for place of articulation in stop consonants.

1. Bursts as invariant place cues

Although numerous investigators have suggested that the burst carries most of the information for place identification, only Cole and Scott (1974a, 1974b) have argued that bursts carry the absolute and invariant cues to place. Their claim for invariance was based on tape splicing experiments using voiced and voiceless stops cross-spliced before three vowels. The results of a much more extensive experiment carried out by Dorman et al. (1977) using digital techniques appear to have adequately refuted the earlier sweeping claims of invariance by Cole and Scott. Dorman et al. studied the identification of /bdg/ before nine VC contexts produced by two talkers. Their results showed that listeners could identify place from bursts spliced onto steady-state vowels segments for less than half of the stimuli. In other words, while there was some invariance, the cues were not <u>sufficient</u> across all vowel contexts.

These previous tape splicing experiments, however, produced unnatural stimuli containing spectral discontinuities at the juncture of the tape splices. The stimuli used in the present experiment provide a better test of the Cole and Scott invariance hypothesis. In this experiment, as well as that of Ohde and Sharf (1977), burst-only segments of natural stop syllables were presented to subjects for identification. The average correct

identification of the burst-only segments for /d/ and /g/ in Experiment 3a was 80%, while in Ohde and Scharf the average was 96.5% (/b/ in Experiment 3 always contained one pitch pulse so it was not included in this average). The apparent discrepancy between these two experiments can be accounted for in terms of the durations of the burst stimuli. Our results showed that performance for place identification improved with longer stimulus durations. Since the syllables in the Ohde and Sharf experiment were read in a list-reading style, the average duration of the bursts for /d/ and /g/ was 29 ms (1977, Table I) compared to 18 ms in Experiment 3. When the average stimulus duration in our experiment was increased to 30 ms by the addition of one pitch pulse, average identification increased to 91%. Furthermore, since the 96.5% average identification for Ohde and Sharf also included bursts for voiceless stops with an average duration of 74 ms, the outcome of both experiments can be seen to be quite similar. Thus it appears that sufficient information for place can be found in isolated bursts, if the bursts are sufficiently long. However, focusing attention on the voiceless burst, obscures the basic problem. Necessary and sufficient information for identification of place depends on the duration of the waveform segment regardless of whether it is aperiodic or periodic.

In summary, although the results of Experiment 2 support Cole and Scott's claim that invariant acoustic cues

may be specified for place of articulation in stop consonants, the results of Experiment 3 do not support the stronger view that the acoustic cues are simple acoustic invariants located in the aperiodic burst of the waveform. Moreover, our results demonstrate that sufficient information for place of articulation is encoded in terms of duration of the acoustic signal extending in time across the fine temporal variation of the stop waveform.

2. Formant transitions as place cues

The second hypothesis in its more general form is that formant transitions themselves contain the acoustic cues which specify place of articulation. This hypothesis as specifically proposed by Liberman (Liberman et al., 1967) assumes that the primary acoustic cue for place of articulation is the direction and extent of the second formant and to some degree the third formant transitions. This hypothesis is based on results obtained from synthetic speech stimuli containing no bursts. Since all three of our experiments addressed some aspect of this hypothesis, we will briefly summarize the major results here. The findings in these experiments were all based on natural stop consonants produced before five or eight different vowels.

In Experiment 1, measurements of formant transitions revealed that the F2 transition could be used to categorize place of articulation before six of the eight vowels examined as long as the vowel context was known. This

result is consistent with Liberman's hypothesis, although the vowel /a/, a popular choice for synthetic stimuli in past studies, was one of the vowels for which F2 was not distinctive. Other parameters of formant transitions in Experiment 1, however, did not provide sufficient information to differentiate place except for a combined measure of F2 X F3 onset frequencies.

Experiment 2 showed that a set of visual features observable in running spectral displays could be used to identify place of articulation independently of the vowel context. Running spectral displays revealed continuous change of spectral prominences from the burst into the voiced formants. However, the specific feature definitions did not make any reference to formant transitions. In fact, in running spectral displays, such as those shown in Fig. 6, no visible distinction was observed between the release burst and voiced formant transition segments for F2 and higher formants. Thus, the formant transitions appear to lose their status as distinct and independent acoustic cues to place in running spectral displays.

Finally, in Experiment 3, the results showed that place of articulation could be identified accurately in the first 20 to 40 ms of a waveform, that is, waveform segments which preceded the actual formant transitions. In fact the acoustic information present in the formant transitions improved the perception of place only slightly beyond information already available in the burst plus one pitch

pulse. That is, formant transitions do not carry the necessary information for specifying place because identification of place is consistently high in waveforms without the transitions.

In addition to these three studies are the results from two perception studies using natural CV's in which the aperiodic segments were removed and only the voiced transitions and following vowel were presented to listeners for identification. Both Dorman et al. (1977) and Ohde and Sharf (1977) reported that place could be accurately identified for only about one-third of the transition-only CV waveforms studied. Thus, for about two-thirds of the syllables, sufficient information for specifying place was not present in the natural formant transitions alone. Taken together these results indicate that necessary and sufficient information for identifying place of articulation in natural stop syllables is not always found in the voiced formant transitions. This conclusion contrasts with the widely held belief that the formant transitions provide the most important information for perception of place (cf. Liberman et al., 1967).

At this point, it seems appropriate to inquire why earlier synthetic speech experiments using only stimuli with differing formant transitions have been so successful in identifying place. If we compare the formant transitions used in synthetic speech stimuli (e.g. Liberman et al., 1967; Harris et al., 1958; or Stevens and Blumstein, 1978)

with the formant transitions measured from natural stimuli in Experiment 1, we can observe a number of differences. For example, the formants for F2 and F3 come artifically close together in synthetic velars compared to the measured values shown in Fig. 3. These differences between the natural and synthetic transitions may provide some additional information which enables listeners to identify place in the transition-only synthetic stops -- information which is not actually present in the formant transitions of natural CV syllables.

3. Complementary role of bursts and formant transitions

The third hypothesis is the familiar one that bursts and formant transitions are both acoustic cues to place, but that for a given vowel context, they provide information in a complementary way. One of the earliest statements of this view was made by Cooper et al. (1952) that "bursts and transitions complement each other in the sense that when one cue is weak, the other is strong" (p. 603). Similar observations have been made over the years by Fischer-Jørgensen (1954), Halle et al. (1957), and more recently by Dorman et al. (1977). Again, the underlying assumption is that the release bursts and formant transitions in stops are independent and separable acoustic entities. Thus research supporting a "cue-trading" hypothesis has generally come from tape splicing experiments which freely interchange burst and transition

segments. The results of experiments using these highly artificial and unnatural stimuli have apparently led to a number of erroneous conclusions. For example, Dorman et al. (1977, p. 116) reported that apical bursts were weak cues before back or rounded vowels. Yet in the present experiment, we demonstrated that listeners could identify apicals before back or rounded vowels from the burst-only segment an average of 90% correct (excluding the 6 ms /do/ burst). Apparently, the abrupt discontinuity of spectral information at the juncture of the tape splice interferes with the correct perception of the consonant which can be identified from information actually available in the burst segment.

Although most results in Experiment 3 indicated that sufficient information for identifying the consonant was located entirely within the first 20 to 40 ms of a waveform, there was one exception to this finding. The identification functions for the consonant and vowel in the syllable /gi/ were reversed from the patterns shown for the other CV's in Experiment 3b (as shown in Figures 14 to 18) and might be seen as supporting the cue-trading hypothesis. However, other evidence suggests that the identification of /gi/ may not be an example of cue-trading at all. Rather, the acoustic cues for specifying phonologically velar place (phonetically mid-palatal) appear to be represented poorly in /gi/. First, note that the vowel /i/ was identified accurately for all 6 syllables in Fig. 14, so the vowel

identification of /i/ in /gi/ is not an exception to this pattern. Apparently, the high compact spectra of the release bursts, particularly with /d/ and /g/, enabled subjects to perceive the following vowel correctly as /i/. It appears, therefore, that the problem in identifying /gi/ is confined to identification of the consonant. But why was the velar not identified accurately in /gi/ even for the longest stimulus? An examination of the literature shows that even carefully produced, natural /gi/ syllables were not identified accurately in several previous experiments. In Dorman et al. (1977), the original /gi/ syllables from both talkers were only identified 60% correct. In LaRiviere et al. (1975), their syllable /ki/ was also identified only 60% correct. [Note that we required the original full-length /gi/ syllable to be identified accurately in the preliminary study leading up to Experiment 3.] The problem appears to be a confusion between /gi/ and /di/. Fully 80% of the erroneous responses for /gi/ in Experiment 3A were alveolars. In Experiment 2, we observed that the mid-frequency peaks for /gi/ were associated with the F3 and F4 resonances. As a consequence, the spectral tilt of the /gi/ burst frame is strongly rising towards the F3-F4 peak. As a consequence, many /gi/ bursts in Experiment 2 were easily confused with /d/ bursts. Although the mid-frequency peaks were still prominent for male talkers, they did not emerge as distinctively in the spectra of the female /gi/'s. Therefore, we may conclude

that the articulation of the palatalized /g/ before /i/ produces spectral cues that are not clearly discriminable from /di/. The validity of this observation gains some support from the well-known historical tendency of languages to change /g/ and /k/ before high front vowels to palatal fricatives or to the alveolar affricates /dʒ/ and /tʃ/ (Bhat, 1978). Thus in English we have only few words beginning with "ge" pronounced with /g/ (like geese),while most words are pronounced as /d / (like gene).

In summary, the major results of both Experiments 2 and 3 are in conflict with the predictions of the cue-trading hypothesis. Specifically, the results demonstrated that sufficient information to specify place of articulation lies entirely within the initial portions of a stop-vowel waveform. The atypical results obtained in both experiments for /gi/ need not be considered as an example of cue-trading, but could be explained in terms of the ambiguous acoustic cues for place associated with the mid-palatal place of articulation in English stops .

4. Integration of burst and transition information

In contrast to the first three hypotheses, the underlying assumption of the fourth hypothesis is that burst and formant transitions are not independent or separable acoustic cues. Both Fant (1960; 1973) and Stevens (1975; Stevens and Blumstein, 1978) have proposed that the acoustic cues for place may be found in the information

contained in the initial 10-30 ms of a stop waveform integrated over the bursts and formant transitions. The emphasis in these accounts has been on the unity of the global information contained in the early portion of the waveform irrespective of the boundary between the aperiodic and voiced portions. Other investigators, including Winitz et al. (1972) and Cole and Scott (1974b) have referred to the integration of burst and transitional cues. However, their strategies for perceptual processing are, in fact, based on the separation of burst and transition cues. Accordingly, the independent burst and transition cues are assumed to be separately extracted from the speech signal and then integrated at some higher level into a phonemic or syllabic percept (Cole and Scott, 1974b, p. 371). The distinction between these two uses of the term integration is an important one because the features described, and the experiments conducted on the basis of these two assumptions are quite different. Indeed, Stevens and Blumstein (1978, p. 1301) specifically rejected the validity of the earlier tape splicing experiments in which conflicting and discontinuous spectral cues were presented to listeners.

Fant and Stevens concur that burst and transition information are not independent precisely because the information arises from a single underlying articulatory gesture. For Fant, the acoustic features for place, as discussed previously, comprise both spectral and temporal properties since the short-time spectra may change either

slowly or rapidly over time. But the spectral prominences (formant patterns if voiced) are always continuous because of the continuity of the underlying stop-vowel articulatory gestures (Fant, 1968, p. 223).

For Stevens, the acoustic features of place are described rather differently. The spectral properties of place are separated from the temporal ones in certain respects. The spectral properties of place are located in a single, static onset spectrum beginning with the burst. These onset spectra, which are assumed to be context-invariant, are considered to be the primary cues to perception of place of articulation. The temporally varying properties of the place cues are located in the formant transitions. These cues are context-dependent and they are considered to be secondary cues to perception of place of articulation (Blumstein and Stevens, 1979, p. 1015). According to Stevens, the primary onset spectrum cues are associated with the underlying articulatory gestures, but the secondary, formant transition cues are not. In Stevens and Blumstein's theory, formant transitions merely "provide the acoustic material that links the transient events at the onset to the slowly varying spectral characteristics of the vowel" (1980, p. 660). This point of view is essentially identical to Cole and Scott (1974b). However, considering Fant's theoretical arguments and the results of our running spectral analysis, Stevens' separation of spectral and temporal properties into primary and secondary

cues appears to be arbitrary and theoretically unjustified. If the acoustic features of place relate directly to an underlying articulatory gesture, and if the essence of a "gesture" is dynamic movement in time, then presumably the associated acoustic features should also incorporate spectral movement in time. In our analysis, we have attempted to specify the direct relation between the underlying articulatory gestures for initial stops and the appropriate acoustic features observable in the continuous changes in spectral energy that are associated with those articulatory movements.

A principled reason exists for the differences between the Fant and Stevens conceptualizations of the role of time in place cues. Fant (1968, p.235) has pointed out that there is no simple, independent definition of the concept of acoustic feature. Rather, acoustic features are defined differently depending on the higher level linguistic units to which they are related. Fant has discussed two types of acoustic features, one correlated with linguistically defined distinctive features, and the other correlated with articulatory phonetic events. Stevens' acoustically invariant features (Stevens and Blumstein, 1980) are clearly oriented toward distinctive features. In the remaining sections below, we will first discuss some of the problems encountered by Stevens and Blumstein in trying to define acoustic features for place of articulation as correlates of distinctive features. Next we will summarize

Fant's theory of phonetically oriented acoustic features, and suggest that our running spectral features are aligned more closely with this category of acoustic features.

5. Acoustic features as correlates of distinctive features

Stevens and Blumstein (1980) have stated that their invariant acoustic properties (or features) are the correlates of "phonetic features" or "phonetic categories" without referring to linguistic categories. However, their phonetic feature category names and descriptions are essentially identical to those of distinctive feature theorists (Jakobson et al., 1952; Chomsky and Halle, 1968; Lieberman, 1970). Fant's comments concerning correlates of distinctive features, which are expanded below, seem to apply equally well to Stevens and Blumstein's features.

Acoustic features that are the acoustic correlates of phonological distinctive features must apply to a broad class of acoustically different phonetic segments across a wide variety of phonetic contexts. The advantage, as Fant pointed out, is basically economic; namely a small set of features applies to a larger class of phonetic segments. For example, Stevens intends his onset spectra to identify place in both stops and nasals in several syllable positions. The disadvantage of this approach is that the acoustic characteristics of the features must be defined in such a way that their description embodies only salient acoustic properties that can be extracted from <u>all</u> of the

segments and contexts to which the distinctive features apply. The necessary compromises in acoustic description make the identification of that feature across a range of phonetic types more difficult. Therefore the onset spectra theory of Stevens and Blumstein should be evaluated in terms of the overall success of the proposed templates for identifying place in all consonant categories and syllabic positions. Thus, Blumstein and Stevens' (1979) claim of 84% correct place identification in initial stops using their static template analysis would certainly be impressive as a distinctive feature correlate, so long as place was identified reasonably well in other phonetic contexts as well. Unfortunately, the results presented by Blumstein and Stevens for place identification of nasals and stops in syllable final position using these same templates were quite different. For example, as summarized in the Discussion section of Experiment 2, in the Blumstein and Stevens' template study (1979), nasals were identified correctly no better than chance level, while stops in final position were identified correctly at the point of closure at only 53%. Apparently, the static onset spectra theory will need to be revised rather substantially in order to specify the acoustic correlates of place as a distinctive feature across all phonetic contexts.

6. Phonetically oriented acoustic features

The other approach to describing the acoustic features for phonetic segments is the one taken by Fant (1968). This view is compatible with the analysis of the running spectra described in Experiment 2. The goal of uncovering and describing acoustic features in this approach is to directly relate the distinctive events in the articulatory phonetic space to a set of salient acoustic features. (See Port (1980) for a discussion of phonetic spaces in language.) According to Fant (1968):

> Such a phonetically oriented approach aims at descriptive power in the first place and a economy of description only in the second place. It should be selective enough to keep apart the main allophones of a phoneme. . .. It should also be closely related to coding strategies for speech synthesis and data reduction schemes in automatic speech recognition. (p.235)

The feature analysis proposed in Experiment 2 is precisely this type of phonetically oriented approach. In Experiment 2 we proposed a set of features to locate the presence of syllable initial stops, and another set to identify place of articulation. These features are claimed to be powerful descriptors for the limited phonetic environment of syllable initial stops, but not necessarily to be descriptors of place of articulation in other syllabic positions. As we will see in the next experiment, the running spectral features can also be used as criteria for synthesizing initial stop consonants, because the appropriate spectral and fine temporal features of the

burst are preserved in the feature definitions. Thus, we see that the contribution of running spectral features as acoustic features for speech perception will be at a lower linguistic level than phonological distinctive features. At this level the features may specify place for only the initial allophones of stops, but they will specify place accurately using known properties of the relation between articulatory shape and radiated sound output.

Since we claim that the running spectral features are phonetically oriented features, we need to say more about the definitions of the associated phonetic events in the listener's phonetic space. The results of Experiment 3 suggest possible definitions of the phonetic events perceived for stop-consonant vowel syllables. The first phonetic event is a separable and independent perceptual unit, the initial stop consonant. Support for an independent initial stop consonant derives from the results that listeners correctly identify stop place from initial waveform portions of a CV syllable -- waveform portions in which they could not reliably identify the vowel segment. We assumed that subjects correctly perceived stop manner as well as place in these waveform portions since subjects never reported difficulty in labeling the stimuli as stops. The underlying articulatory gesture for this short initial stop is the release from vocal-tract occlusion at a specified place of articulation. The running spectral features by this account are the acoustic properties which

directly link this articulatory gesture to the perception of the phonetic event. The features are seen to specify the initial stop consonant as a unified set of acoustic properties. They do not necessarily have an independent status as do acoustic features which are supposed to be correlates of phonological distinctive features.

Although the present experiment was not specifically designed as a study of vowel perception, our results do bear on the description of vowels as phonetic events as well. Accurate identification of non-diphthongized vowels was obtained in this study from acoustic information present in the first three pitch pulses of the syllable. Since three pitch pulses encompass the formant transitions for a CV syllable with virtually no steady-state vowel portion, one may conclude that listeners identified the vowel solely on the basis of dynamic information contained in the rapidly moving formant transitions. These findings are consistent with the earlier conclusions of Strange et al. (1976) that dynamic formant transitions carry substantial acoustic information that can be used for vowel identification. Apparently the definition of a vowel as a phonetic event should include a dynamic component. This dynamic component, if present, would have as its underlying articulatory gesture a movement toward a particular vocal-tract shape or target. This is not to deny that there is probably a steady-state component to both vowel articulation and its associated phonetic percept. These

proposals are only meant to stress that there is also a dynamic or time-varying component, one which apparently contains sufficient acoustic information for the identification of non-diphthongized vowels.

The account of stop-vowel perception outlined above in terms of the set of acoustic features that underlie phonetic events is compatible with the recent theoretical position taken by Studdert-Kennedy (1980). Studdert-Kennedy emphasizes that speech perception can only be understood by relating acoustic cues to the underlying motor gestures. He suggests that knowledge of the relation between isolated and discrete acoustic cues and segmental phonetic units has not led researchers to a deep understanding of the speech perception process. However, this understanding can be reached, he argues, by defining the relation between the acoustic signal and the gestures that shape the vocal tract. The set of spectrally changing features examined in Experiments 2 and 3 can be related directly to the underlying gestures that comprise the release of vocal-tract occlusion. In the spirit of the theoretical positions set forth by Studdert-Kennedy and Fant, these proposed features can be seen as a contribution to a phonetic description of initial stop consonants. Moreover, these particular feature definitions were motivated by acoustic analysis and perceptual tests to establish their validity across two independent domains.

We should note here that Studdert-Kennedy's approach (1980) as well as the approach advocated earlier in terms of running spectral features are in agreement with certain premises expressed by Liberman (Liberman et al., 1967). In particular, Liberman has also argued that the acoustic correlates of speech are linked to the dynamic underlying articulatory gestures and, therefore, that the associated spectral cues must incorporate change over time as well. Although Liberman's hypothesis that formant transitions carry the dynamic information for place of articulation in stops has not been supported by our research, the necessity for specifying dynamic change in spectral cues for stop consonants has been fully supported by both Experiments 2 and 3.

V. EXPERIMENT 4. PERCEPTION OF STATIC VERSUS SPECTRALLY CHANGING CUES FOR PLACE IN SYNTHETIC CV'S

A. Introduction

The major results of Experiment 3 indicated that information for identifying place of articulation in stop consonants can be found in the initial portions of a CV waveform. For the consonants /b/ and /d/, presentation of stimuli containing only the first 20 ms of the waveform provided listeners with sufficient acoustic information to identify the stop correctly on 94% of the test trials. For /g/, however, longer waveform durations (at least 40 ms) were needed to reach a level of performance better than 90% correct identification.

We have already examined two different approaches for describing invariant cues for place of articulation in these short waveform portions. The first method, described in Experiment 2, used spectrally changing features to identify place in running spectral displays. The second method, proposed by Stevens and Blumstein (1978; Blumstein and Stevens, 1979; Blumstein and Stevens, 1980), used static cues represented by spectral templates to identify place in a single integrated onset spectrum. The essential difference between these two analysis methods may be characterized in terms of an emphasis on <u>static</u> versus <u>dynamic</u> spectral cues for identifying place. The present experiment was designed specifically to examine differences

in perception between static and dynamic cues to place of articulation in synthetic stop consonants.

The methodology underlying the logic behind Experiment 4 was, in fact, alluded to by Stevens and Blumstein themselves:

> A stronger test of (their) theory would be to determine whether perception of place of articulation depends on attributes of the gross shape of the spectrum at onset, independent of fine details such as burst characteristics and formant onset frequencies (1978, p.1367).

Thus, an experiment which compares the identification of synthetic speech stimuli preserving only the onset spectrum shape with other synthetic stimuli that preserve the fine temporal details of the stop consonant waveforms should provide an important test of the Stevens and Blumstein onset spectrum theory. In the case of the feature analysis proposed earlier in Experiment 2, the fine details of the stop waveform are, in fact, preserved in the running spectral displays and explicitly used to identify place in the feature definitions we developed. For example, the Tilt of burst feature requires identifying a spectral section encompassing only 5 ms of the release burst that is specifically separated from the following aspiration or voiced formant information. Thus, an experiment comparing the identification of a set of synthesized CV's patterned after Stevens and Blumstein's onset spectrum with the identification of synthesized CV's patterned after the running spectral features of Experiment 2 would accomplish

two things. First, the identification performance from the Stevens and Blumstein stimuli could demonstrate whether sufficient place information is actually contained in the overall shape of the onset spectrum, the very test Stevens and Blumstein suggested but did not carry out. Second, a comparison of identification performance between the two sets of synthetic stimuli should indicate whether the underlying static or dynamic cues are a better representation of the critical features that listeners use to identify place of articulation in stop consonants.

The basic strategy of Experiment 4, then, was to examine the identification of short waveform portions synthesized from two sets of rules, one patterned after Stevens and Blumstein and the other based on running spectral features. Since naive subjects were successful in identifying place in the truncated natural speech stimuli of Experiment 3, the procedures developed for this experiment were closely patterned after those in the previous experiment. First, subjects participated in an identification task using truncated natural speech CV's. Subjects who could identify the natural speech stimuli at relatively high levels of performance were then required to identify synthesized stimuli from both the Stevens and Blumstein and the running spectra stimuli sets.

In addition to collecting subjects' identification responses for each stimulus, a confidence rating response was also obtained. A three interval confidence rating scale

was included to allow a subject to indicate whether his response represented a chance guess, whether he was very sure his response was correct, or whether he wished to indicate a confidence rating between these two choices. The purpose of collecting confidence ratings was to provide more information about a listener's perception than the results obtained from forced choice identification. For example, we were interested in determining whether subjects would be as confident about the correct identification of the synthesized stimuli as they were about the natural stimuli. Thus, the confidence ratings provide an additional measure by which to compare the natural and synthetic speech stimuli.

The success of Experiment 4 depends on clearly stated and executed principles for synthesizing the Stevens and Blumstein (hereafter called S+B) and the running spectra (hereafter called RS) stimuli. Each synthesized syllable was specifically modeled after the appropriate natural syllable using visual spectral matching techniques. Fortunately, the tools for doing extremely good modeling were available. The spectral analysis of the natural stimuli was carried out by linear prediction analysis as implemented in the SPECTRUM program (see Appendix A). The S+B and RS stimuli were synthesized on the KLATT digital synthesizer (Klatt, 1980a) as implemented in the KLTEXC program (Kewley-Port, 1978).

The variables manipulated in this experiment were selected in accordance with the results obtained in the previous experiment. Three consonants /b,d,g/ were paired with the three vowels /i,a,u/ to produce a base set of nine CV syllables. Since identification performance varied with the durations of the stop waveforms presented in Experiment 3, waveform duration was also selected as an experimental variable. Three durations of the truncated stops were selected, 20, 30 and 40 ms. Based on the results of Experiment 3, these durations spanned the 20 ms duration at which /b/ and /d/ were accurately identified, to the 40 ms duration at which /g/ (except for /gi/) was accurately identified. Thus, four independent variables were manipulated in this experiment, stimulus type, consonant type, vowel type, and stimulus duration.

The overall goal of the synthesis strategy was to keep as many parameters as possible the same between the S+B and RS sets, while at the same time incorporating the important static versus dynamic spectral differences implied by the two place cue theories. A general description of the criteria used to synthesize the S+B and RS stimuli is summarized below, while the detailed descriptions are found in the Procedure section.

According to Stevens and Blumstein, the important acoustic information for place of articulation is located in the 25.6 ms onset spectrum. Using the KLATT synthesizer, it was possible to generate a steady-state stimulus such

that its spectrum at any point matched the onset spectrum of the original natural CV. However, it was not apparent whether these steady-state stimuli would be perceived as stop-consonants or vowels. In the speech perception literature, it is generally assumed that a rising F1 transition is an important manner cue for the class of stop consonants (Delattre et al., 1955; Stevens and House, 1956; Fant, 1960, Liberman et al., 1967). For this reason, Blumstein and Stevens (1980, p. 651) used rising F1's in their synthesis of otherwise steady-state consonant stimuli.

The issue of the presence of a F1 transition as a cue to stop manner became problematic for this study when the actual measured formant transitions of Experiment 1 were considered. In Experiment 1, we found that the measured F1 transitions for the vowels /i/ and /u/ (two of the vowels to be used in this study) were actually flat or falling (see Fig. 3). Since naturally occurring CV syllables that have flat F1 transitions are still perceived as stops, it was not clear what kind of F1 transitions should be used in the synthetic S+B stimuli. Therefore, a pilot experiment was first carried out using S+B stimuli synthesized with and without F1 transitions. The purpose of this experiment was to determine if listeners would judge the F1 transition stimuli as more stop-like than the F1 steady-state stimuli. The details of this experiment are presented below in Section C. Based on the outcome of the pilot experiment,

measured F1 transitions from the natural CV's were used as synthesis parameters in the S+B stimuli.

For the RS stimuli the underlying synthesis principle was to preserve the spectral place cues implied by each of the three spectral features defined in Experiment 2. The Tilt of burst feature was preserved by carefully matching the first spectral frame of the RS and natural speech stimuli. The feature of Late onset of F1 was preserved by using the average VOT values measured in Experiment 1. For the /g/ syllables, the Mid-frequency peak feature was synthesized by matching the voiceless spectral sections between the RS and natural stimuli.

Thus, the stimuli for this experiment consisted of three sets of brief, truncated stop-vowel syllables. One set consisted of naturally produced CV's, and the other two sets were synthesized after models of the static versus dynamic acoustic cues for place of articulation. Subjects were required to identify the initial consonant, and judge how confident they were in the correctness of their response. The results from this experiment should help evaluate the validity of the claims made by Stevens and Blumstein that static onset spectra can serve as perceptual cues for specifying place of articulation in stop consonants.

B. Stimuli: Synthesis parameters

Two sets of synthetic syllables were carefully modeled after spectra calculated from a set of natural speech syllables. The first task was to select an appropriate set of natural CV syllables. The natural stimuli consisted of the nine base syllables spoken by talker RP. Each syllable had to be carefully analyzed by the SPECTRUM program to determine that place of articulation could be correctly identified by both the Blumstein and Stevens' templates and the running spectral feature analysis.

The SPECTRUM program calculated the onset spectrum in this study by a procedure identical to that used by Blumstein and Stevens (1979). The first 25.6 ms of the CV waveform was windowed using a unique window shape invented by Blumstein and Stevens. The first 12.6 ms was rectangular, followed by a 12.6 ms one-half Hamming window which may be thought of as an extended half-Hamming window. Blumstein and Stevens (1979) actually used an extended half-Kaiser window to calculate their spectra. A Kaiser window is similar to a Hamming window but it is considered to be an improvement because peak to valley differences in the resulting spectrum are greater (Rabiner and Gold, 1975). Differences in the linear prediction spectra resulting from an extended half-Hamming window compared to an extended half-Kaiser window are small, however, and would probably not change the shape of any of their

proposed templates. In any case, the synthesis modeling techniques used in this experiment always employed the same extended half-Hamming window. The onset spectrum was then calculated exactly as Blumstein and Stevens calculated it. The waveform was preemphasized, windowed and analyzed with 14 linear prediction coefficients using the autocorrelation method. A 17 inch refresh CRT display (DEC VT-11) allowed for direct visual comparisons between onset spectra for the natural and S+B synthesized syllables.

Each of the nine syllables from Experiment 3 was examined according to the following criteria using the SPECTRUM program. The experimenter visually examined all onset spectra and only accepted tokens whose spectra clearly fit the overall template descriptions of diffuse-rising, diffuse-falling and compact and also appeared to meet all the template rules described in the Blumstein and Stevens study (1979). The running spectral display for each syllable was examined to see that it contained good exemplars of the running spectral features. An additional criterion for velar syllables was also checked. The onset spectrum for these syllables had to include an F1 peak so that the resulting synthesized waveforms would all have a voiced component. If all criteria were not met, another token was drawn from one of the five tokens examined in Experiment 1. In the end, five natural CV's, /bi,di,ba,bu,gu/ were taken from the stimuli used in Experiment 3, and the rest were substituted from the original tokens obtained in Experiment 1.

The nine natural syllables were then edited digitally to approximate the 20, 30 and 40 ms waveform durations as closely as possible. The stimuli were edited at zero crossings at the end of pitch pulses. The average durations of the 20, 30 and 40 ms natural stimuli were 21 ms, 30 ms and 39 ms respectively.

The synthesis of the S+B and RS sets was done with the KLATT synthesizer configured as a parallel formant synthesizer with six formants (Klatt, 1980a). Glottal resonance characteristics were shaped for talker RP's /i/ and then kept constant. Nasal resonances were not used. The fundamental frequency was set to a constant 100 Hz. Synthesis parameters always terminated exactly at the 20, 30 or 40 ms durations. Fortunately, no synthesized waveforms had appreciable baseline offsets since the fundamental was 100 Hz.

Insert Figure 19 about here

The exact procedure for generating an S+B stimulus was as follows. Figure 19 shows the match between the onset spectra of the final synthesized S+B stimulus for /ga/ and the natural /ga/ stimulus. SPECTRUM calculated the onset spectrum of the natural stimulus and then printed out the frequency, bandwidth and relative amplitude values for each spectral peak. These values were then used as input synthesis parameters for F1 to F6 in the KLATT synthesizer.

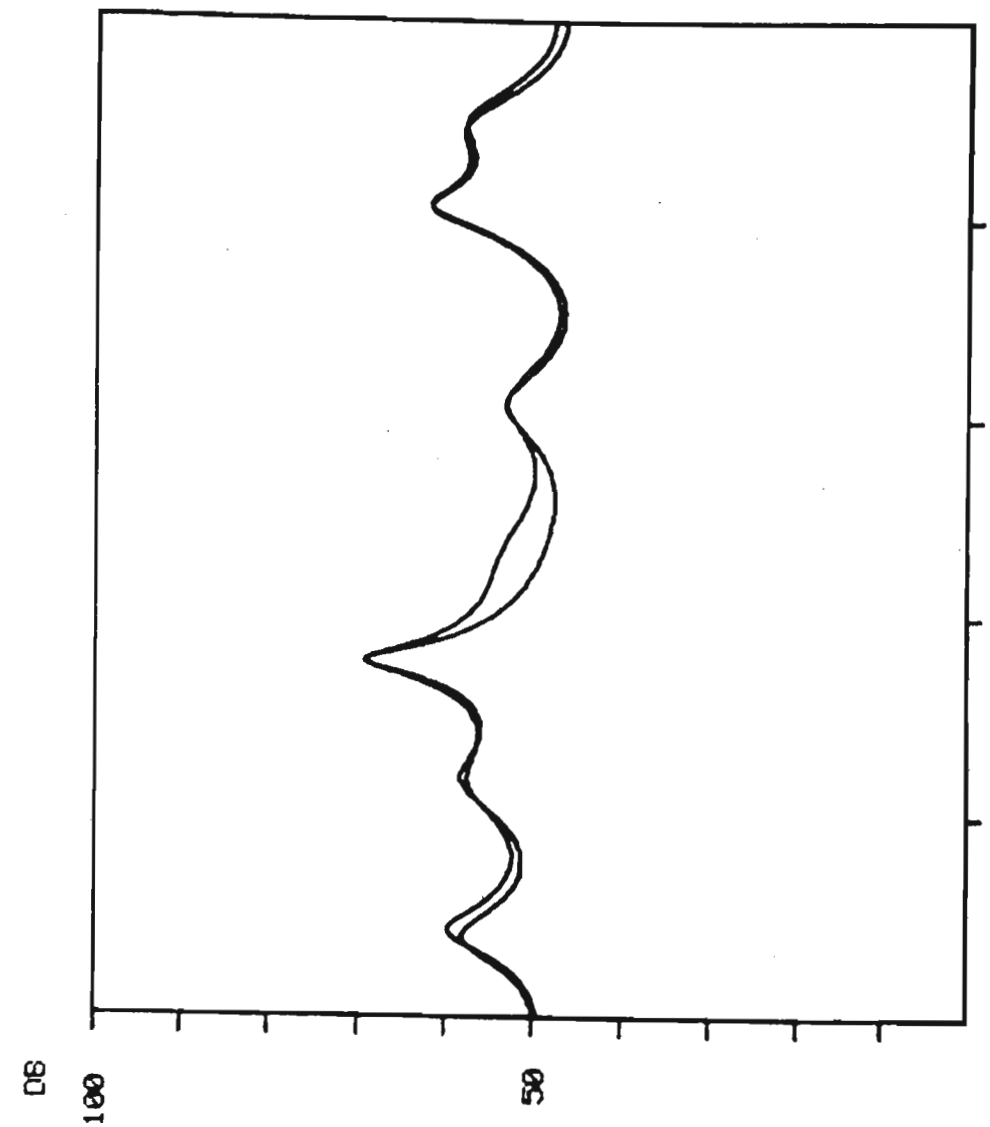


Figure 19. Comparison of the 25.6 ms onset spectra for natural speech /ga/ and S+B synthetic /ga/.

A good spectral match was obtained by resynthesizing and adjusting only the bandwidth and amplitude parameters. The voicing source (AV) was always set to its maximum value.

Two sets of S+B stimuli were generated for use in the pilot study of the effects of F1 transitions on the perception of stop manner. In the first set, all synthesis parameters were held constant. The second set of S+B stimuli was generated by substituting the steady-state F1 parameter values with the measured F1 transition values obtained from the natural stimuli. The onset spectra for the two sets of stimuli were indistinguishable. Both sets of S+B stimuli were generated at 20, 30 and 40 ms durations.

Synthesis parameters for the RS stimuli were derived from several sources of information. First, the running spectrum was calculated according to the procedures described in Experiment 2. Second, the average formant transition values measured in Experiment 1 were obtained from tables given in Appendix B. VOT values were chosen by approximating the average VOT values measured in Experiment 1 as summarized in Table 4. VOT was kept constant across place, /b/ = 0 ms, /d/ = 10 ms, and /g/ = 20 ms. The amplitude of the voicing source always started at a value that was 5 dB less than the maximum, and increased to the maximum in 5 ms. Using the measurements from Experiment 1, steady-state formant frequencies for the first four formants were determined for the vowels /i/, /a/ and /u/.

The transition parameters of frequency and duration appropriate for each CV were then inserted after the release burst.

For /b/ and /d/, the shape of the first frame containing the burst was carefully matched to the natural stimuli on the CRT. The first frame in these running spectral displays had an effective window of 5 ms, compared to 20 ms for the Stevens and Blumstein onset spectrum. After the burst frame, the formant transitions previously inserted gave the shape of the succeeding RS frames. Amplitudes of the formants in the parallel synthesizer were adjusted to match the amplitudes observed in the running spectra of the natural stimuli.

Procedures for synthesizing the mid-frequency peaks for /g/ differed somewhat from those for /b/ and /d/. Figure 20 shows the running spectrum for the natural /ga/, and the final synthesized RS /ga/. In order to preserve the mid-frequency peaks shown in the voiceless frames, careful spectral matching for each of these frames was often needed. The synthesis matching procedure focused on the spectral tilt of the burst frame, and on the mid-frequency peaks feature of the voiceless frames, but not on other spectral properties of the voiceless frames. Thus the overall procedure for generating the RS stimuli was not based on a frame-by-frame spectral match of the natural and synthetic stimuli. While the burst frames were spectrally matched, which meant one frame for /b/ and /d/, and the

three voiceless frames for /g/, the remaining synthesis parameters were determined essentially by rule using values from the averages of the formant transitions found in Appendix B.

Insert Figure 20 about here

When a running spectral match was achieved visually according to the above criteria, we listened to the synthetic waveform. All stimuli sounded like good exemplars of the intended syllable, and none of the synthesis parameters were adjusted on the basis of listening. Therefore, whatever merits or weaknesses were incorporated into the actual synthesis of the S+B and RS stimuli, the procedures for both sets were developed on the basis of visual spectral matching alone. The final RS stimuli were then synthesized at the 20, 30 and 40 ms durations by deleting parameter lines in the formant transitions which extended beyond the specified values.

C. Stimuli: Role of F1 transitions

Stevens and Blumstein have claimed that a rising F1 transition is a necessary cue for the perception of stimuli as stop consonants (Blumstein and Stevens, 1980; Stevens and Blumstein, 1980). In developing the S+B stimuli for Experiment 4, two goals were kept in mind. On the one hand, the S+B stimuli were intended to have steady-state

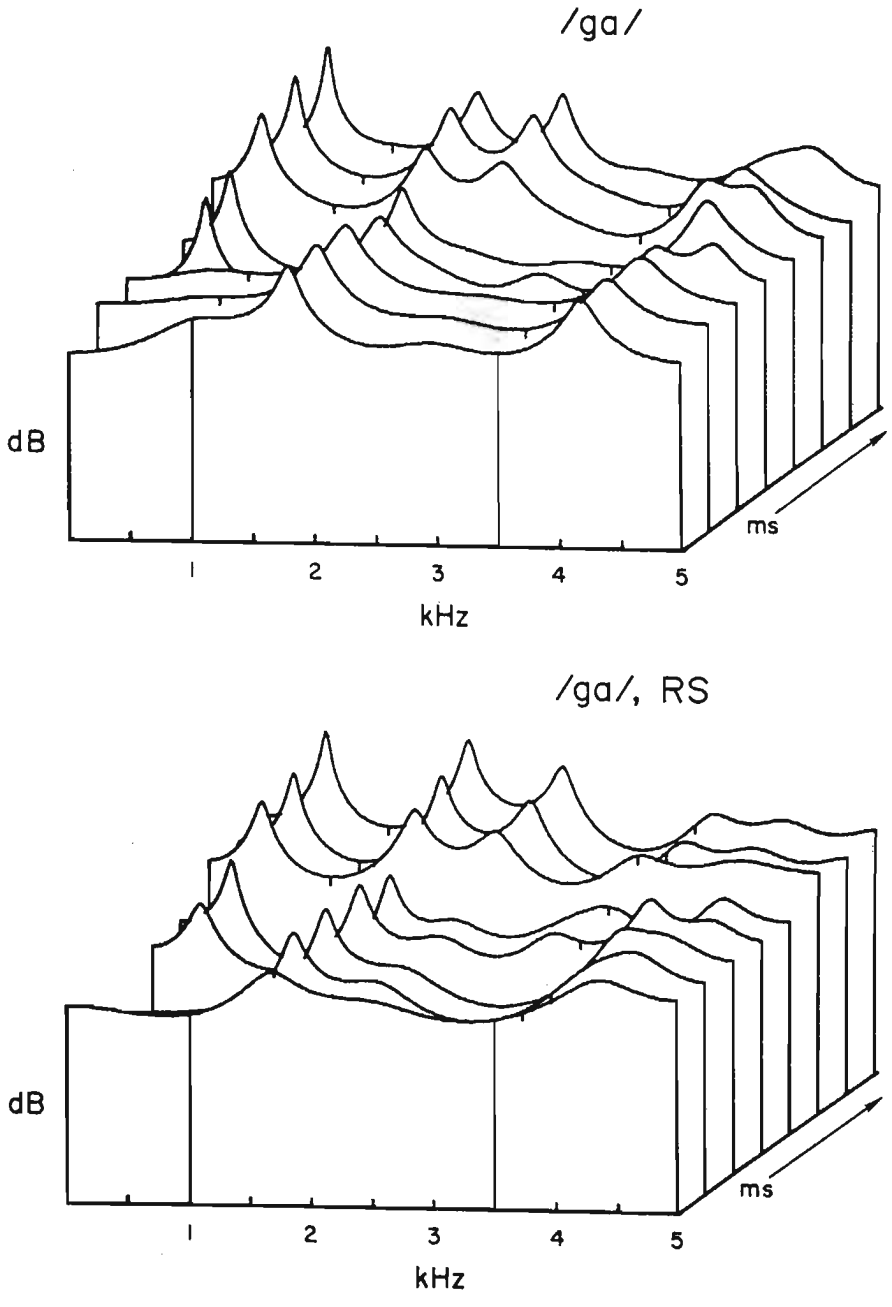


Figure 20. Comparison of the running spectral displays for the natural speech /ga/ and the RS synthetic /ga/.

resonance properties encoding only the matched onset spectrum. On the other hand, the purpose of the experiment was to examine place cues for stop consonants, and therefore the stimuli were supposed to be as "stop-like" as possible. To determine if the S+B stimuli with F1 transitions were more stop-like than the F1 steady-state stimuli, we carried out the following pilot experiment.

The purpose of the pilot experiment was to determine whether listeners would identify the 40 ms stimuli in these two sets as the consonants /b,d,g/, or as some indeterminate vowel. Four listeners trained in phonetics served as subjects. A computer program randomized 10 repetitions of the nine stimuli in each set and output them through the D/A converter every three seconds. Subjects listened to the stimuli in a quiet testing room over TDH-39 headphones. They were instructed to try to identify the sounds as /b,d,g/, but if the sounds seemed not to have a stop-like quality they were to respond with V for a vowel-like sound. No training or familiarization procedures preceded testing.

The responses were scored in two ways. First, the number of consonant responses for each stimulus was tallied. Second, the overall correct consonant identification for each stimulus was calculated. In order to determine whether the S+B set with the F1 transitions were identified better by either of these two measures, the percentage difference between the two stimulus sets also was calculated.

The results showed that the identification of the steady-state S+B stimuli as consonants ranged from 100% to 22%. The /bi/, /bu/ and /du/ stimuli were always identified as consonants, while /gi/ and /ga/ were identified as consonants for only about one-quarter of the responses. The average was 63%. We therefore conclude that a rising F1 transition is not a necessary cue to stop consonant manner for all syllable types, a finding that is contrary to the earlier claims by Blumstein and Stevens (1980, p. 660).

Nonetheless, it is possible that the presence of a F1 transition might improve the stop-like quality of a steady-state stimulus which was frequently identified as a vowel, e.g. /gi/ and /ga/. An examination of these results showed that for stop consonants before the vowels /i/ and /u/, only a slight (3%) improvement in stop-like quality was obtained for the set with the rising F1 transition. For the vowel /a/, however, /ba/ and /da/ showed substantial (30%) improvement in stop-like quality over the steady-state F1 stimuli. The /ga/ stimulus, however, showed the reverse, namely 12% more vowel-like responses. Thus, an overall 7% improvement in stop-like responses was obtained for the set of stimuli with the F1 transition, although the source of the improvement came from only 2 of the 9 syllable types, /ba/ and /da/. Also, it should be noted that the three stimuli, /gi/, /ga/ and /gu/, which displayed the poorest consonant identification, showed no improvement in number of consonant responses with the F1

stimuli. Results from the other measure showed only a 6% improvement in the correct identification of the initial consonant with the S+B stimuli containing F1 transitions. The source of the improvement was again the /ba/ and /da/ stimuli which improved an average 35% correct. Average correct consonant identification was actually slightly worse for the /i/ and /a/ stimuli with F1 transitions than for the steady-state stimuli.

In summary, the results of the pilot experiment showed that F1 transitions were not universally necessary for the perception of short, speech-like stimuli as stop consonants. However, two of the nine S+B syllables with F1 transitions showed some improvement in stop-like quality over the steady-state stimuli. Therefore, the S+B stimuli with the F1 transitions were selected as the more "stop-like" of the two sets of S+B stimuli originally developed and they were used in the main experiment reported below. They will henceforth be referred to as the S+B stimulus set.

D. Procedure

The testing procedures used in this experiment were controlled on-line by a PDP-11/05 computer. Stimuli were output at 10 kHz sampling rate through 12-bit D/A converters, low-pass filtered at 4.8 kHz and then presented over TDH-39 earphones at approximately 85 dB SPL. Button press responses were collected from up to six subjects at a

time and stored on disk. Manipulation of cue lights and feedback lights was also under computer control. (See Forshee, 1979, for a more detailed description of this system.) The basic design of the present experiment was intended to follow closely that of Experiment 3, allowing for the appropriate changes from tape recordings and written responses to an on-line computer controlled perceptual experiment.

The testing procedures were divided into a two day sequence. Day 1 served to screen subjects audiometrically, and to train and test them with the natural speech stimuli. On Day 2, subjects identified only the synthetic stimuli. The separation of the natural speech and synthetic stimuli experiments served several purposes. Based on Experiment 3, it was apparent that subjects needed some familiarization with these short 20 to 40 ms speech stimuli in order to identify them reliably as consonants. In this experiment we decided that the familiarization tasks would be best carried out with only the natural speech stimuli. Subjects therefore listened to the synthetic stimuli on Day 2 with strategies derived entirely from their exposure to natural speech stimuli on Day 1.

Another result obtained in Experiment 3 was that not all subjects correctly identified the consonant in the longer stimuli. In particular, some subjects consistently identified /bi/ as alveolar or /gi/ as bilabial. In this experiment, we decided that the comparison between the

three sets of stimuli would be facilitated, if all of the natural speech syllables were identified at fairly high levels of performance. Therefore, a criterion was developed for selecting subjects to participate in the synthetic stop conditions of the experiment on Day 2 based on the results from Day 1. This criterion required subjects to identify each of the nine natural speech syllables, averaged over all durations, at better than a 40% correct level. Since 33% correct is chance, this criterion effectively excluded only those subjects who consistently misidentified some syllable type and subjects who were performing essentially at chance.

Responses were collected on a seven button response box. At the top of the response box was a cue light which was illuminated briefly just before the presentation of each test stimulus. The seven buttons were laid out in one row from left to right. Above each button was a feedback light. The three buttons on the left were labeled for consonant identification. As in Experiment 3, the goal of the experiment was to separate place cue responses, bilabial, alveolar and velar. Voicing was not an experimental variable. Therefore, the consonant response buttons were labeled "B/P" for bilabial, "D/T" for alveolar, and "G/K" for velar. The lights above the labeled response buttons also served as feedback lights in the familiarization tasks.

On the right side of the button box were three additional buttons labeled with the confidence rating responses. Confidence ratings were only collected in the identification parts of this experiment. The confidence rating categories were defined as follows on the written instruction sheets:

Confidence Ratings:

++ Very sure the consonant was correctly identified.

+ Reasonably sure the consonant was correctly identified.

- Consonant response represents only a chance guess.

The three buttons were labeled above with -, +, ++ and below with GUESS, SURE, VERY SURE from left to right. On each trial, after an identification button was depressed for a given stimulus, the three lights over the confidence rating buttons were illuminated to indicate that a confidence response was to be entered. These lights remained on until either the subjects pressed a rating button or a three second response period elapsed. After both an identification and a confidence rating response were collected for each stimulus, a new test trial was initiated.

Insert Table 12 about here

Testing procedures are summarized in Table 12. Each subject was screened by an audiometric test for the octave frequencies from 500 to 8000 Hz at a sound pressure level of 20 dB (ANSI-1969) using a Grason-Stadler Model 1701 audiometer. Two familiarization tasks were conducted using the 27 truncated natural speech stimuli. The first, called "cued familiarization," required listening only. All 27 natural stimuli were presented once each in order from the longest to shortest durations. Before each stimulus was presented, the light over the correct response button was illuminated.

The second familiarization task was more similar to the forced-choice identification task. Two repetitions of each natural stimulus were randomized and output over the headphones. A cue light signaled the onset of each stimulus. After hearing the stimulus, subjects pressed a response button. The correct response light was then illuminated to give subjects appropriate feedback. Confidence ratings were not collected during familiarization.

Subjects were then instructed on how to do the identification with confidence ratings task. For the natural set of stimuli, subjects heard five blocks of two repetitions of each stimulus. A consonant identification

Table 12. Procedures for the consonant identification with confidence rating task in Experiment 4.

Day 1: Natural speech stimuli

Task	Stimuli
1. Screening audiometry	
2. Cued familiarization	27 natural, ordered
3. Feedback familiarization	54 natural, randomized
4. ID with confidence ratings	270 natural, randomized (5 blocks)

Day 2: Synthetic RS and S+B stimuli

Task	Stimuli
1. ID with confidence ratings	540 RS and S+B, randomized (10 blocks)

and rating response were collected for each test stimulus. Stimuli were fully randomized, and stimulus presentation was paced to the slowest subject's responses. Overall, ten consonant and rating responses were collected for each natural stimulus from each subject. No feedback was given in this task.

After the Day 1 testing was completed, a computer program analyzed the subjects' identification responses. The number of correct responses averaged over stimulus duration was calculated for the nine natural syllables and printed out separately for each subject. Any subject who scored less than 40% correct on any of the nine syllables was telephoned and asked not to return for the second day of the experiment. Subjects were told when signing up for the experiment that they might participate in one or two days of testing depending on their performance on Day 1.

Subjects returning for Day 2 were given instructions explaining that the identification and confidence rating procedures were identical to the procedures used on Day 1. They were not told about differences in the nature of the stimuli to be presented. Instructions merely stated, "You will be listening to additional short consonant stimuli one-at-a-time." No familiarization trials were presented on Day 2 and subjects began the identification with confidence ratings task directly, having previously listened to only the natural speech stimuli.

The 27 S+B stimuli and the 27 RS stimuli were fully randomized within one stimulus block of 54 trials. Stimulus presentation was paced to the slowest subjects' responses. Subjects listened to ten blocks of stimuli, with a 30 second pause between blocks, and a five minute break half-way through the experiment. Overall, ten consonant and confidence rating responses were collected for each of the 54 synthetic stimuli presented to each subject.

Subjects were obtained through a laboratory paid subject pool. None of the subjects who participated in Experiment 3 were contacted for this experiment. Subjects were paid $3 a day for testing. Subjects were advised in advance that a screening audiometric test would be administered before testing.

E. Results

Twenty-one subjects were tested in Experiment 4. One subject did not pass the screening audiometric test and was dropped from further testing. Ten subjects did not achieve the 40% correct level of performance on the natural speech stimuli on Day 1 and were asked not to return for testing on Day 2. Thus, data were collected from all three sets of stimuli for 10 subjects, resulting in 100 data points per stimulus.

Insert Table 13 about here

A summary of the results obtained for the three stimulus sets is shown in Table 13. The first row presents the percent correct consonant identification for each stimulus set averaged over the consonant and vowel types, and stimulus duration. As in Experiment 3, subjects showed considerable accuracy in identifying place from very brief initial portions of natural stop consonants. The overall percent correct was 94%. Performance levels for both sets of synthetic stimuli fell below that of the natural speech stimuli. Nonetheless, consonants were identified better in the RS stimuli with 78% correct than in the S+B stimuli with 68% correct.

The results from the consonant identification task, were confirmed by analysis of variance which showed that the natural, RS and S+B stimuli formed three distinct ordered categories of stimulus type ($F(2, 8) = 78.28$, $P < .001$); a one-way analysis of variance with Scheffé post hoc analysis at the $p < .05$ level revealed three distinct stimulus groups. Consonants were identified more accurately for the natural speech stimuli than for the RS synthetic stimuli, and more accurately for the RS stimuli than for the S+B synthetic stimuli.

The next three rows in Table 13 summarize the results of the confidence rating responses. Results were computed

Table 13. Percentage of response measures collected in Experiment 4 presented for each stimulus type averaged over vowel and consonant type, and stimulus duration.

Response measures	Natural speech	Synthetic, RS	Synthetic, S+B
Correct consonant identification	94	78	68
Guess rating (-)	4	3	11
Sure rating (+)	21	21	26
Very sure rating (++)	75	76	63

as the percentage of the number of each confidence rating response obtained for each stimulus set. These results show clearly that subjects used the same confidence ratings for their responses to both the natural and RS stimuli. However, subjects were far less confident of their responses for the S+B stimuli. For example, subjects rated only 3% to 4% of their responses as guesses for the natural and RS stimuli, but 11% as guesses for the SB stimuli. These results suggest that there were some aspects of the natural and RS stimuli that subjects judged as similar to each other, and dissimilar to the S+B stimuli.

Insert Figure 21 about here

The main effects of each of the experimental variables -- consonant type, vowel type and duration on the percent correct consonant identification for each stimulus set are shown seen in Fig. 21. The main result of decreasing performance levels of identification from the natural to RS to the S+B stimuli was preserved across all variable types but one. Identification of /b/ syllables for the S+B set was better than the RS set, although identification for the natural /b/ was still better. The main effect of stimulus set was perfectly preserved for the vowel and duration variables shown in Fig. 21.

The effects of each of the experimental variables on consonant identification were then examined in more detail.

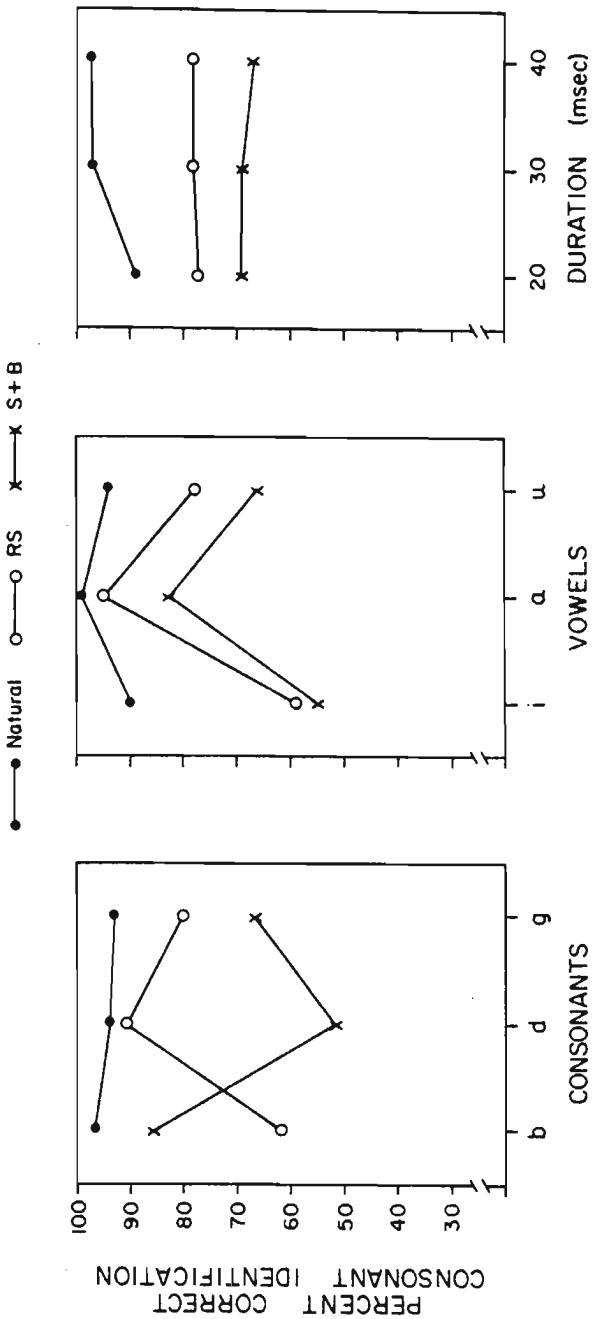


Figure 21. Percent correct consonant identification shown separately for each stimulus type, natural, RS and S+B. Each panel displays results for one independent variable averaged over the other two independent variables.

For consonant types, no differences in identification performance among /b/, /d/ and /g/ were observed ($F(2, 8) = 1.05$, NS). As shown in Fig. 21, differences in performance level were obtained for the vowels ($F(2, 8) = 67.63$, $p < .001$). Consonants were identified better (93% correct) before the vowel /a/. Consonants before /u/ were identified better (80% correct) than consonants before /i/ (68% correct). Furthermore, a one-way analysis of variance using a Scheffé post hoc analysis showed that consonant identification performance formed three distinct groups for vowel type. Vowel context, therefore, had an important effect on the identification of consonants in the short stimuli used in this study. It is interesting that the vowel /a/, which has often been used in synthetic speech perception studies involving stop-vowel syllables, showed the most robust vowel context effect and was the easiest for subjects to correctly identify in the synthetic stimuli.

No overall difference in consonant identification was observed across the 20, 30 and 40 ms waveform durations ($F(2, 8) = .88$, NS). A slight improvement between the 20 ms and 30 ms durations was observed for the natural speech stimuli. As shown below, however, this improvement was contributed by only three of the nine syllables.

The results showing percent correct consonant identification are broken down separately by each experimental variable in the nine panels on Fig. 22. This

figure illustrates the major sources of the variation observed in the average identification functions shown on Fig. 21.

Insert Figure 22 about here

The ordered performance levels of the natural, RS and S+B stimuli were preserved for all variable types except the consonant /b/. Figure 22 shows that consonant identification was better for the S+B compared to the RS sets for /bi/ and /bu/, but not /ba/. (Surprisingly, /ba/ had 100% correct identification for all stimulus sets, for all subjects!) A post-hoc spectral analysis of the RS /bi/ and /bu/ stimuli revealed something previously overlooked. Figure 23 shows the running spectra of the natural /bi/ stimulus in the top panel and the RS /bi/ synthesized stimulus in the bottom panel. Although the spectral tilt and shape of the burst frames were carefully matched, the relative levels of signal energy between the burst frame and the vowel differed between the two stimuli. That is, for /bi/, RMS energy in the sixth frame for both stimuli was identical. For the natural stimulus, the burst was 19 dB lower than frame six, but for the RS stimulus, the burst was only 13 dB lower. Therefore, the burst for RS /bi/ was 6 dB louder (relative to the vowel) than the burst for the NS /bi/. Similarly, the burst of the RS /bu/ was 6 dB louder (relative to the vowel) than the burst for the NS /bu/.

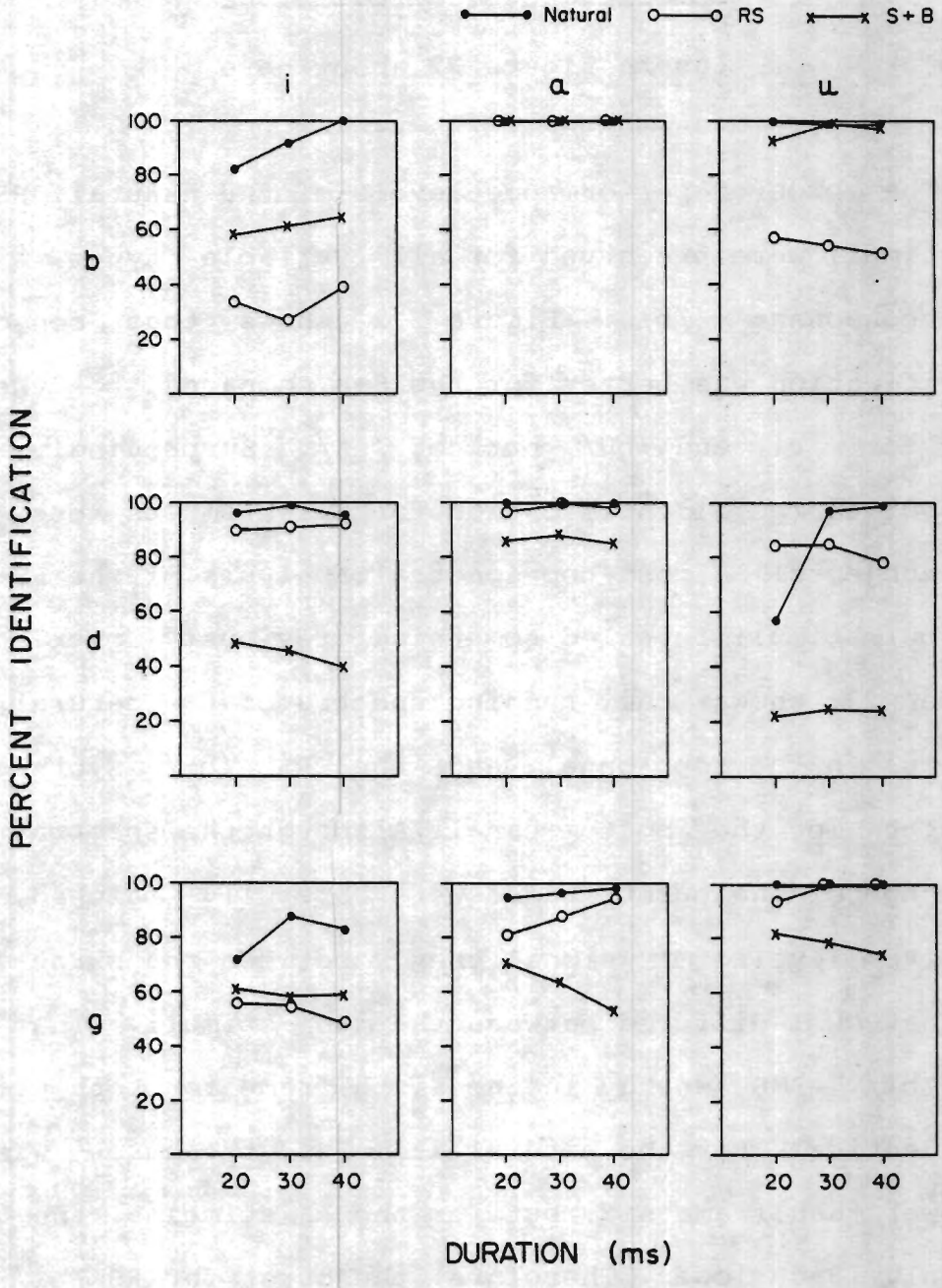


Figure 22. Percent correct consonant identification, plotted separately for each experimental variable.

Insert Figure 23 about here

The higher amplitude synthesized bursts were a direct result of the RS synthesis rule which implemented a constant voicing source for all synthesized RS syllables. This rule, which was intended to keep the source constant across place, was in some sense a mistake. It is well-known that labial bursts have substantially weaker signal energy than alveolar or velar bursts. For example, Zue (1976) measured the spectral amplitude of stop bursts relative to the vowel for 15 vowels and found alveolar and velar burst amplitudes to be approximately equal, but labial bursts were 12 dB lower. In the case of the synthetic RS /bi/ and /bu/ stimuli, we do not know for certain that place of articulation was misidentified on the basis of the unnaturally loud bursts. But the possibility is there. It would appear from the results of this synthesis study that burst energy may serve as an important perceptual cue to place of articulation and therefore should be examined more carefully in future research.

When the overall effects of the consonant and vowel type of syllable on the consonant identification were discussed above, differences were found for vowel context, but not for consonant type. If we examine the results shown in Fig. 22 in greater detail, we can see that the effect of vowel context was not uniform across the consonants. In

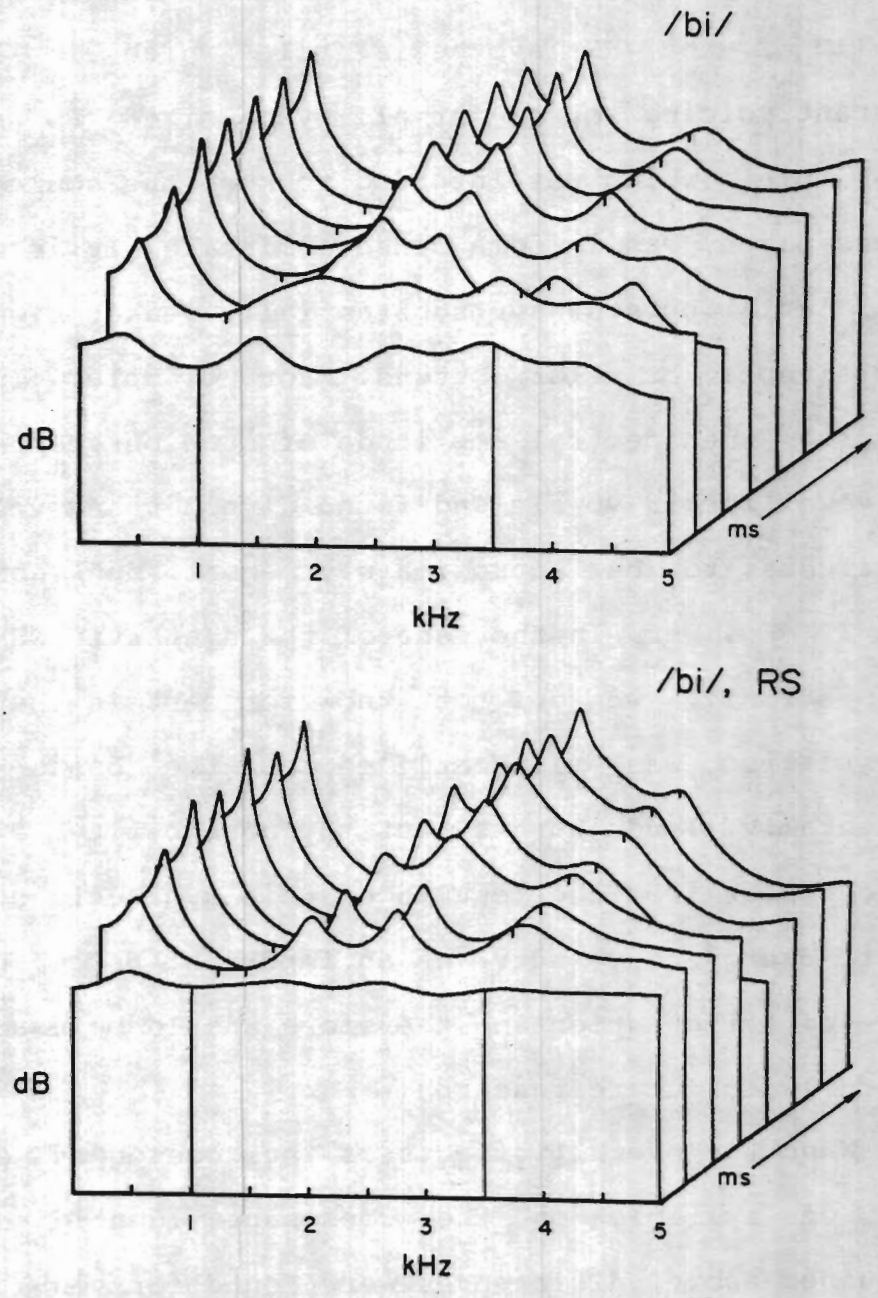


Figure 23. Comparison of the running spectral displays for the natural speech /bi/ and the RS synthetic /bi/.

fact, an interaction between consonant and vowel types was observed ($F(4, 24) = 26.22$, $p < .001$). In addition, interactions between consonant type, vowel type and stimulus set were also observed (interaction of stimulus X consonant, $F(4, 24) = 36.63$, $p < .001$; interaction of stimulus X vowel, $F(4, 24) = 7.07$, $p < .001$; interaction of stimulus X consonant X vowel, $F(8, 32) = 10.77$, $p < .001$). These results indicate that the individual consonant-vowel syllables contributed differentially to the overall effects of the ordered difference in performance level between the natural, RS and S+B stimuli.

As shown in Fig. 22, the 20, 30 and 40 ms waveform durations had only a very small effect on identification performance. Based on the results obtained in Experiment 3, we expected that several natural speech stimuli especially /gi/, would show an improvement in consonant identification with stimulus duration. This result was observed here since /bi/, /da/ and /gi/ all showed increased identification performance with longer duration waveforms. For the synthetic RS and S+B stimuli, however, consonant identification did not improve with longer stimulus durations. Overall, then, the results of varying stimulus duration in this experiment affected consonant identification only minimally. The 20 ms difference between the longest and shortest stimuli was probably not large enough to reveal significant perceptual effects.

Confidence ratings were also collected in this experiment to provide additional information concerning perceptual differences between the three stimulus sets. If the three category confidence rating scale was a reliable measure in this study, then the number of correct responses should be systematically associated with higher confidence ratings. This relation was examined by calculating the conditional probability of obtaining a correct response, C, for each rating category, Rj, as P(C/Rj). Conditional probabilities for each stimulus set, for each rating category were calculated separately and the results are plotted in Fig. 24.

Insert Figure 24 about here

Subjects' confidence rating responses were highly correlated with their ability to identify place of articulation correctly in the three sets of stimuli. The conditional probabilities displayed here are, of course, a combined measure of the correct identification of place and the selected confidence rating category. (These measures are also reported separately in Table 13.) This combined measure, displayed in Fig. 24 also shows the ordered effects of performance on the three stimulus sets, from natural to RS to S+B. The natural speech stimuli clearly had higher conditional probabilities than either of the two synthetic stimulus sets. However, for all confidence rating

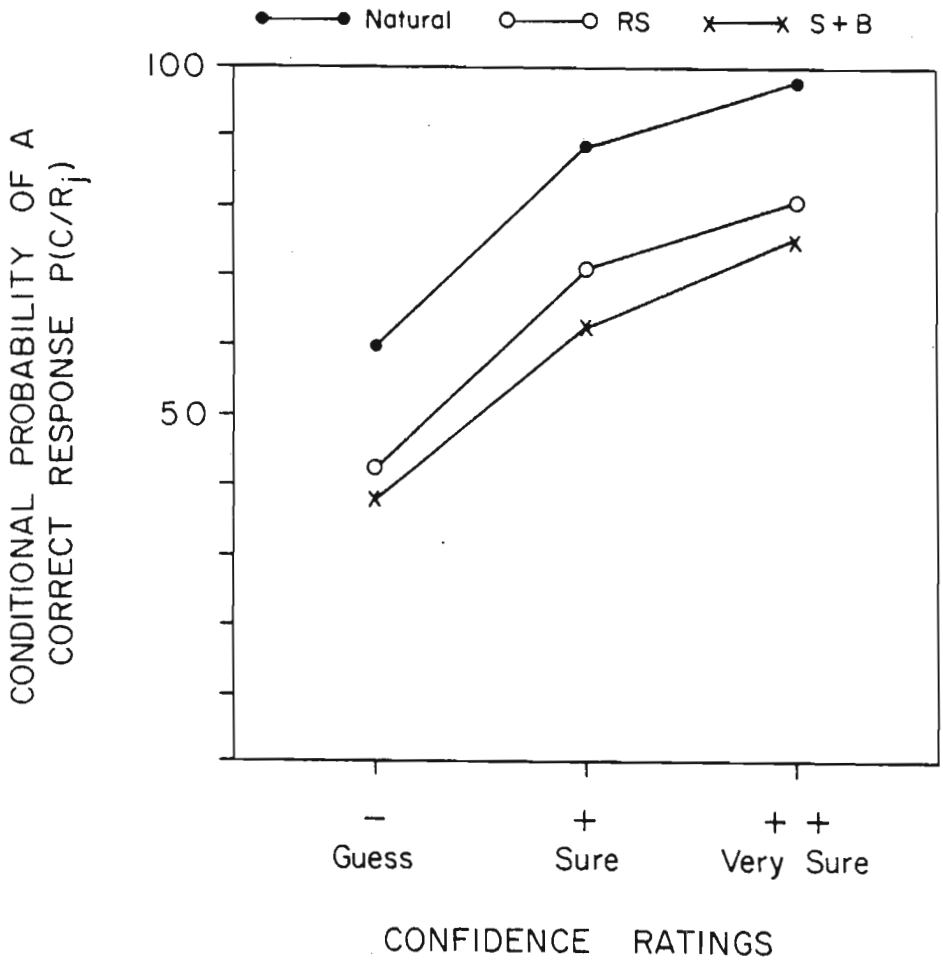


Figure 24. Conditional probability of a correct response obtained for each confidence rating plotted separately by stimulus type, natural, RS and S+B averaged over consonant type, vowel type, and duration.

categories, subjects were more likely to identify place of articulation correctly from the RS stimuli than the S+B stimuli. Thus the conditional probabilities showed both the reliability of the confidence rating responses, and the overall main effect of stimulus type found in the earlier results based on only percent correct consonant identification.

F. DISCUSSION

The present experiment examined the perception of place of articulation in natural and synthetic stop-vowel syllables. Two sets of stimuli, natural and RS, contained dynamically changing spectral information while the S+B set contained static information representing only the overall gross shape of the onset spectra as described by Stevens and Blumstein. Identification performance of the S+B stimuli was intended to provide an empirical test of the validity of Stevens and Blumsteins onset spectra theory for perception of place of articulation. The results showed that listeners were only able to correctly identify place in 68% of the S+B stimulus trials. This performance level was 26% worse than the data obtained from the same listeners for comparable natural speech stimuli. Of the nine S+B syllable types examined, only three, /ba/, /bu/ and /da/, were identified better that 85% correct. The remaining six CV's averaged only 55% correct identification. These perceptual results indicate that

subjects cannot reliably identify place of articulation from only the overall gross shape of the onset spectra of these short stop CV waveforms. Apparently the information in the onset spectra alone is not sufficient to serve as reliable and sufficient cues to place of articulation in stop consonants.

This conclusion is strengthened by the comparison between the S+B stimuli and the RS stimuli. The RS stimuli were synthesized using rules that preserved the dynamically changing spectral features for place that were derived in Experiment 2. The RS stimuli were modeled after the same natural speech stimuli as the S+B stimuli. As a consequence, the calculated onset spectra of an RS stimulus should also be a good exemplar of the appropriate onset spectra for a particular place of articulation. A post hoc comparison of the onset spectra for the RS stimuli and the S+B stimuli showed the RS onset spectra were good place exemplars. (In fact, for half of the comparisons the onset spectra for the RS and S+B stimuli were virtually indistinguishable.) Thus, both the static onset spectra properties and the dynamic running spectral features are present in the RS stimuli. However, identification results showed that place was correctly identified for 78% of the short RS stimuli. Thus, subjects listening to two sets of synthetic stimuli, both with the appropriate onset spectra, showed a 10% increase in identification for RS stimuli which contained dynamically changing place information in

addition to the static onset spectra. Ten percent may not represent a large overall increase, but the sources of the increase as shown in Fig. 22 demonstrate important differences between the two stimulus sets. For the three S+B syllables which had high levels of identification, two of the RS stimuli, /ba/ and /da/, also had high levels of identification. But for the six S+B stimuli with poor identification, four of the RS stimuli showed substantial improvement (/di,du,ga,gu/). In fact, the average identification scores for these six stimuli increased from 55% for the S+B set to 75% for the RS set. Thus the addition of dynamic temporal information to the static onset spectra in synthetic CV's resulted in a significant increase in place identification.

Can this experiment also be considered a test of the validity of the individual running spectral features as place cues in initial stops? Since the experiment was not designed with this goal in mind, the results in fact do not bear directly on the evaluation of the individual features. The running spectral features were defined as contrastive binary features. An appropriate test of these features would be to manipulate them independently in synthetic stimuli and observe the perceptual differences. For example, the bursts in the RS /ba/ and /da/ could be altered parametrically along a Tilt of burst continuum from rising to falling. Listeners judgements of the /ba/ to rising-/ba/ and /da/ to falling-/da/ continua could

establish the effectiveness of the Tilt of burst feature in place identification.

As noted above, this experiment uncovered a previously overlooked acoustic feature to place of articulation. We discovered that the /bi/ and /bu/ stimuli were poorly identified as bilabials apparently because the signal energy of the burst was too high relative to the vowel. Synthetic speech experiments, such as the one suggested above should be able to determine the role of burst energy as an additional acoustic cue for place of articulation. Of course, differences in the actual acoustic signal energy among the release bursts for /b/, /d/ and /g/ must be established empirically as well.

On the other hand, the RS stimuli were identified correctly on 87% of all trials except for the stimuli /bi/ and /bu/. Furthermore, for half of the CV's, the identification functions for the natural speech and RS stimuli were virtually indistinguishable as shown in Fig. 22. The synthesis procedures for the RS stimuli were primarily rule governed and not based on frame-by-frame spectral matching. Given this strategy, the modest difference between 87% for the RS stimuli and 94% for natural stimuli suggests that the rules used in the synthesis were quite adequate in providing the appropriate acoustic information for specifying place. Therefore, the results of this experiment support the hypothesis that the individual running spectral features provide reliable and

sufficient acoustic information to specify place of articulation in stop consonants.

Confidence rating responses were collected from subjects in addition to the consonant identification responses. The conditional probabilities shown in Fig. 24 demonstrated that subjects could assign the confidence rating responses in a consistent and reliable manner. The results in Table 13 also established that subjects were equally confident about identification of the natural and RS stimuli, but were far less confident of their responses for the S+B stimuli. These differences in confidence ratings can be accounted for by a simple hypothesis. Subjects were probably responding to underlying structural differences in the stimulus waveforms. Both the natural speech and RS stimuli preserved the fine temporal details of the stop waveform and therefore had dynamic changes in their spectra over time. The S+B stimuli, on the other hand, were essentially static signals with little change in the fine temporal structure over time. The grouping of the confidence rating responses reflected these underlying dynamic versus static differences. We conclude that our subjects were sensitive to the dynamic versus static differences in these stimuli sets, even though the durations of the stimuli were extremely brief, 20 to 40 ms.

The overall results of this experiment call into question the entire approach of Stevens and Blumstein to invariant place cues for stop consonants. First, we found

that the gross shape of the spectrum at onset did not provide sufficient cues to place of articulation. This finding was demonstrated very clearly in a synthetic speech experiment that followed an earlier suggestion of Stevens and Blumstein. Second, we found that there was a substantial increase in place identification from the static S+B to the RS synthetic stimuli which had preserved the underlying dynamic structure of the acoustic signal for stops. Moreover, subjects showed by their confidence ratings that they were far more confident of responses to short stimuli with the appropriate dynamic temporal variations than they were of static S+B stimuli. Thus subjects are quite sensitive to dynamic spectral information in initial stops, and they can make consistent use of this information to correctly identify place from even synthetic versions of CV syllables that preserve this important acoustic information. Taken together, the results indicate that the static onset spectra approach of Stevens and Blumstein does not provide sufficient acoustic information to reliably specify distinctive perceptual cues for place of articulation in stop consonants.

In summary, the results of Experiment 4 establish that dynamically changing spectral information is a better representation of the acoustic cues for place in initial stop consonants the than the static information contained the gross shape of the spectrum at onset. The running spectral features from Experiment 2 were used to synthesize

dynamically changing CV waveforms that preserved the fine temporal structure of natural speech. These dynamic features appear to be promising as reliable and sufficient acoustic cues for the perception of place of articulation in stops.

VI. SUMMARY AND CONCLUSIONS

The major conclusion we wish to draw from this investigation is that invariant acoustic cues for specifying place of articulation in initial stop consonants can be found in the first 20 to 40 ms of the stop waveform. In a general sense, our findings are in agreement with Fant (1973) and Stevens and Blumstein (1978; Blumstein and Stevens, 1979). However, we claim that descriptions of these acoustic cues must include both spectral and temporal properties in order to capture the relevant temporal differences associated with the underlying articulatory gestures (Fant, 1960; Fant, 1973). This conclusion is also in general agreement with the views of Liberman and other investigators at Haskins Laboratories who have emphasized the dynamic nature of the speech cues (Liberman et al., 1967; Dorman et al., 1977; Studdert-Kennedy, 1980).

In Experiment 1 we showed that the spectrally changing information present in voiced formant transitions was not sufficient for distinguishing place of articulation among stops across a number of vowel contexts. Therefore, following a suggestion of Searle et al. (1979), running spectral displays were created in Experiment 2 to examine spectrally changing information from the release burst continuously into the voiced formant transitions. Known spectral and temporal properties associated with the stop consonant release gestures were used to define a set of

three visual features. We then conducted an experiment to determine if these features would be descriptively adequate for specifying place of articulation for initial stops in several vowel contexts produced by three talkers. The results showed that the visual features were invariant cues across vowel context, but not across differences in vocal tract size. Post hoc analyses suggested, however, that simple rules could be incorporated in the feature definitions to account for differences in vocal tract size. Critical-band (auditory) filtering was implemented in the analysis procedures to make the displays comparable to the output from the human auditory system. All three acoustic features for specifying place were found to be displayed robustly in the auditory filter running spectra.

Two speech perception experiments were conducted to confirm the observations made in Experiment 2 and to evaluate the earlier claims of Stevens and Blumstein (1978; Blumstein and Stevens, 1979; Blumstein and Stevens, 1980). In Experiment 3 short initial portions of stop-vowel waveforms from two talkers were presented to observers for identification of either the consonant or the vowel. For the consonants /b/ and /d/, identification was about 95% correct with only 20 ms of the initial waveform. The velar consonant /g/ required a longer duration waveform, 30 to 40 ms, for accurate identification, although /gi/ was not identified accurately in segments even as long as 90 ms. These results suggest that sufficient acoustic information

for identifying place of articulation in stops resides in the first 20 to 40 ms of a stop waveform.

The important difference between the acoustic cues proposed by Stevens and Blumstein (1978) and the running spectral features developed in Experiment 2 was an emphasis on the temporal dimension. The properties of the onset spectra proposed by Stevens and Blumstein have no time dimension, whereas the running spectral features proposed here incorporate specific changes in spectral energy over time. In Experiment 2 we argued that the running spectral features were an improvement over the static onset spectra because the temporal characteristics of the articulatory gestures and resulting acoustic signal are in fact saliant properties for distinguishing place of articulation in stop consonants. These temporal characteristics distinguished /g/ from /b/ and /d/. Results from Experiment 3 were more compatible with our analysis in terms of running spectral features, incorporating temporal characteristics, than with an analysis in terms of the static onset spectra proposed by Stevens and Blumstein.

A set of synthetic CV stimuli was constructed in Experiment 4 to preserve only the static onset spectral information in the acoustic waveform. Results from identification tests showed that the cues contained in the onset spectra alone did not provide reliable and sufficient acoustic information to specify place of articulation. Moreover, results obtained from similar synthetic CV

stimuli which preserved the temporal structure of the stop waveforms showed significantly higher levels of identification performance. These synthetic stimuli were constructed to preserve the running spectral features developed in Experiment 2. A comparison of the identification of these synthetic stimuli with appropriate natural CV stimuli indicated that the running spectral features provide sufficient acoustic information for specifying place of articulation in initial stops.

In conclusion, this investigation has demonstrated that a set of spectral and temporal properties can be defined sufficient to distinguish place of articulation in stops over several vowel contexts and speakers. The temporal properties of stops, which are particularly important for identification of velar consonants, cannot be captured in static displays of the overall gross shape of the onset spectrum as Stevens and Blumstein (1978) have proposed recently. Both the spectral and temporal properties associated with different places of articulation in stops are very distinctive in linear prediction running spectra. In this investigation we have shown that a small number of features can be defined to distinguish place of articulation in running spectral displays of natural speech. Moreover, we have shown that these same features can serve as salient perceptual cues in synthetic CV syllables. Thus, change in spectral energy over time appears to be an appropriate characterization of the way

that the peripheral auditory system encodes speech signals. We conclude, therefore, that spectral change should be an integral part of descriptions of the acoustic cues that listeners use to perceive place of articulation in stop consonants.

APPENDIX A. SPECTRUM: A Program for Analyzing the Spectral Properties of Speech

A. Introduction

The SPECTRUM program was developed in the Speech Perception Laboratory at Indiana University, by the author, to be a general purpose laboratory program used in the spectral analysis of speech signals. SPECTRUM was designed to permit flexible spectral analysis using interactive computer graphics for user feedback. SPECTRUM has been under continuous development throughout the course of this research. Many of the major findings of this dissertation are to a large part the result of the author having direct control over subroutines developed for SPECTRUM.

The primary analytic techniques employed in SPECTRUM were based on linear prediction analysis, but other types of routines have also been implemented in the package including discrete Fourier transforms and spectral analysis using a digital model of an auditory or psychophysical filter bank. To a large extent, SPECTRUM was patterned after the ILS (Interactive Laboratory Systems) programs developed at Speech Communication Research Laboratories, Los Angeles, Ca. Most of the signal processing algorithms can be found in Markel and Gray's (1976) book, Linear Prediction of Speech, except for the FFT which was developed by Markel (1971).

SPECTRUM was written in FORTRAN IV to operate on small laboratory computers. SPECTRUM was originally developed to

run on a PDP 11/05 with a DEC VT-11, a seventeen inch refresh memory scope for interactive graphics. The scope and package of programs were later transferred to a faster PDP 11/34 machine which has now become the major device used for signal processing in the laboratory. SPECTRUM will run in 28K of core under the DEC RT-11 operating system. One 2.6 megabyte disk drive was dedicated to data storage. A 12 bit A/D and two D/A converters were used to digitize or listen to the speech waveforms. Further details of the hardware are described by Forshee (1979).

B. Human Engineering

In developing a highly technical program like SPECTRUM, one must carfully consider the type of user and the most frequent program applications. In this case, most laboratory users are non-engineers and have speech research interests involving relatively small data bases. The research applications usually involve questions concerning the spectral properties of speech which can be analyzed best using on-line, interactive graphics. It was therefore decided that a high priority goal was easy man-machine communication, even when this resulted occasionally in less flexible data manipulation.

To this end, the analysis file structures were fixed and made opaque to the user. SPECTRUM automatically saves almost all calculated values, such as linear prediction coefficients, formant values etc., in the integer-valued

analysis file without the user specificaly designating which output values should be saved. One rationale for this type of analysis file was that calculation time is relatively costly, at least on the 11/05 machine and the small data bases are often examined repeatedly in detail.

Another important programming feature was to base the analysis interval on the concept of the frame as developed in ILS. The frame is defined by the number of waveform samples between succesive spectral analyses of a waveform segment. In order to time-lock any given analysis file to a waveform file, the frame becomes the important defining characteristic of the analysis file. This gives the user an advantage of being able to analyze frames out of sequence, or recalculate frames using different analysis conditions. Once the context has been set for a particular analysis file, it cannot be changed, although several different analysis files having different contexts can be associated with a single waveform file.

The SPECTRUM commands are selected by means of two character names listed in the MENU shown in Table A-1. After a command is

---

Insert Table A-1 about here

Table A-1. SPECTRUM MENU of commands available in August, 1979.

2 Character Command	Description
AC	Compute Analysis Coefficients
CS	Display Cursor VT-11
CT	Query or change Context
EX	Exit Spectrum
FT	Compute and display Formant Tracks
FR	Redisplay FT
GP	Query and change Global Parameter
GT	Grid for TD display
HC	Hard copy from Tektronix
LI	List on printer
MU	Menu display
OW	Open Waveform and analysis files
OA	Open new analysis file
PP	Pick Peaks spectrum and display
PR	Redisplay PP
QD	Query Length Display Buffer
QP	Query pointers
SI	Sift Pitch Extractor
SW	Redisplay SI
SS	Calculate Smoothed Spectrum and display
SR	Redisplay SS
TD	Three-dimensional spectrum display
TR	Redisplay TD
TE	Transfer to Tektronix and erase first
TN	Transfer to Tektronix but not erase
TH	Transfer, erase and hardcopy
TI	Title VT-11 display
XF	Calculate Fast Fourier spectrum and display
XR	Redisplay XF
WA	Display Waveform VT-11
WR	Redisplay WA

accessed by SPECTRUM, a prompt message is displayed on the CRT screen. In some cases another menu may also be displayed. These prompts contain mnemonics separated by commas to indicate the input arguments for a specific command. For example, command PP (i.e., pick the peaks from a spectral section) has the prompt:

STFR,NO.FR,ERASE->

'STFR' means frame where the analysis is to start; 'NO.FR' means the number of frames included in the analysis; and 'ERASE' is a binary valued parameter to erase (or not erase) the graphics terminal before the PP output is displayed.

The structure of these prompts has been carefully designed. First, the prompts include all arguments used in the command. The order of the arguments, as well as any abbreviations used, is preserved across commands. Second, input values for the arguments operate in a full default mode so that reasonable values for all arguments are assumed by SPECTRUM. (The default mode will be discussed further below.) Detailed documentation of command actions and arguments (similiar to the ILS format) are available to the user. An example of the documentation for command SS is found at the end of this Appendix.

C. Program Structure

The program SPECTRUM consists of a root segment, a subroutine library, and a set of command subroutines with essentially one subroutine per command. The root contains the FORTRAN COMMON blocks and the calling sequences for each command. The individual command subroutines are then used in overlay regions of core. Subroutines used by several commands, such as FFT, reside in the library.

The full default mode of the commands depends on passing argument values in COMMON. The two most important command arguments in COMMON are STFR and NO.FR. Once set by one command, they are retrieved from COMMON by all subsequent commands until they are reset. Another use of COMMON is to maintain the global analysis parameters which allow continuity of analysis conditions during processing. Parameters such as M (i.e., number of linear prediction coefficients), or NBITS (i.e., power of the FFT) are kept in a single global table. As global parameters are set for specific analyses, they can then be recalled in subsequent analyses from the COMMON. Furthermore, they are stored in the header of the analysis file so that the most recent global parameters are reset in COMMON anytime an analysis file is opened.

There are generally two types of output produced by a command. Typically, any values calculated are stored in the analysis file, and appropriate displays of these values are

plotted on the graphics terminal. The listing command, LI, permits a listing of all stored values on either the line printer or the user's console. The displays are currently plotted on the DEC VT-11. Displays are not usually flexible, but rather oriented to a particular task, as can be seen in the next section. Because the displays are constructed relatively slowly, the most recent display from a command is stored on disk for rapid retrieval (see the MENU). Originally, hard-copy of the displays was obtained using a Polaroid CU-5 camera attached to a custom-made hood for the VT-11. Currently a Tektronics 4010 terminal and storage oscilloscope with a Tektronics 4631 hard copy unit are used to reproduce the displays on sheets of 8 X 11 inch silver oxide paper.

D. Sample Protocol with Displays

In order to illustrate the use of SPECTRUM, sample protocols for research tasks similar to those used in this dissertation are outlined below. Also included are references to the specific algorithms incorporated in SPECTRUM.

The first task is to analyze in detail the early formant transitions in a consonant-vowel syllable. A waveform file and analysis file are opened for the syllable /go/. The frame size is set to 50 points or 5ms. The waveform is displayed for the first 30 frames (150 ms) using command WA as shown in Figure A-1.

Insert Figure A-1 about here

The first 100 ms or 20 frames of a syllable were always analyzed. The command AC calculates the linear prediction coefficients as inverse filter coefficients using the autocorrelation method. This analysis uses the subroutine AUTO from Markel and Gray (1976). Command PP is executed next to calculate both the smoothed spectral sections for each frame using subroutine FFTMGR from Markel, (1971), and to pick the spectral peaks for each frame using subroutine FINDPK from Markel & Gray (1976). The resulting display of the peaks and RMS energy for each frame is shown in Figure A-2.

Insert Figures A-2 & A-3 about here

The formant tracks were calculated for the first four formants of /go/ by FT, subroutine FORMNT fom Markel & Gray (1976). The display of the 20 frames of waveform, the peaks and the superimposed formants are shown in Figure A-3.

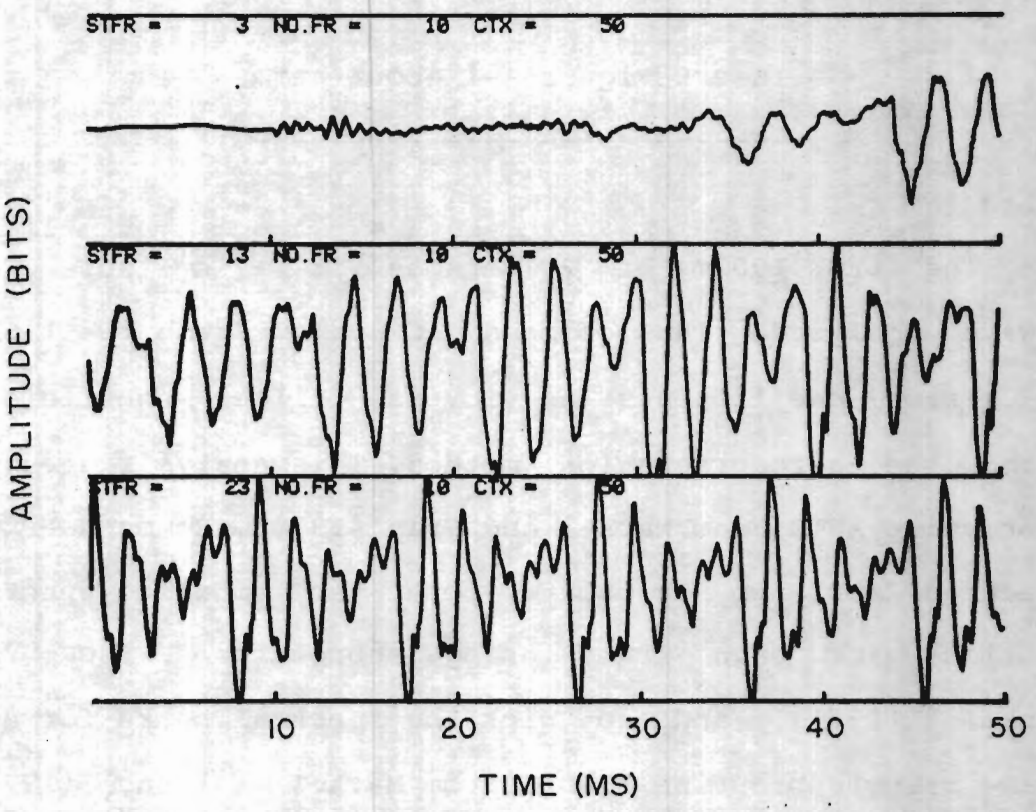


Figure A-1. The first 30 frames of /go/ are displayed sequentially at three different locations on the CRT screen using command WA.

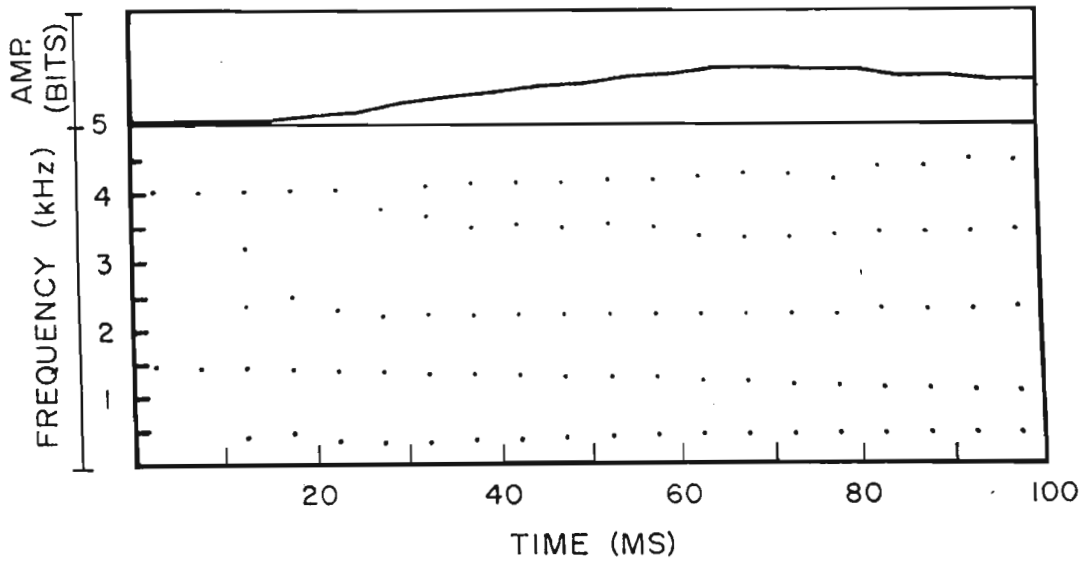


Figure A-2. The display of the spectral peaks for each of 20 frames produced by the command PP, plus the RMS energy curve. The x-axis is time in frames. The y-axis ticks for the formant peaks are at 500 Hz intervals. The y-axis range for RMS energy is 0 to 32767.

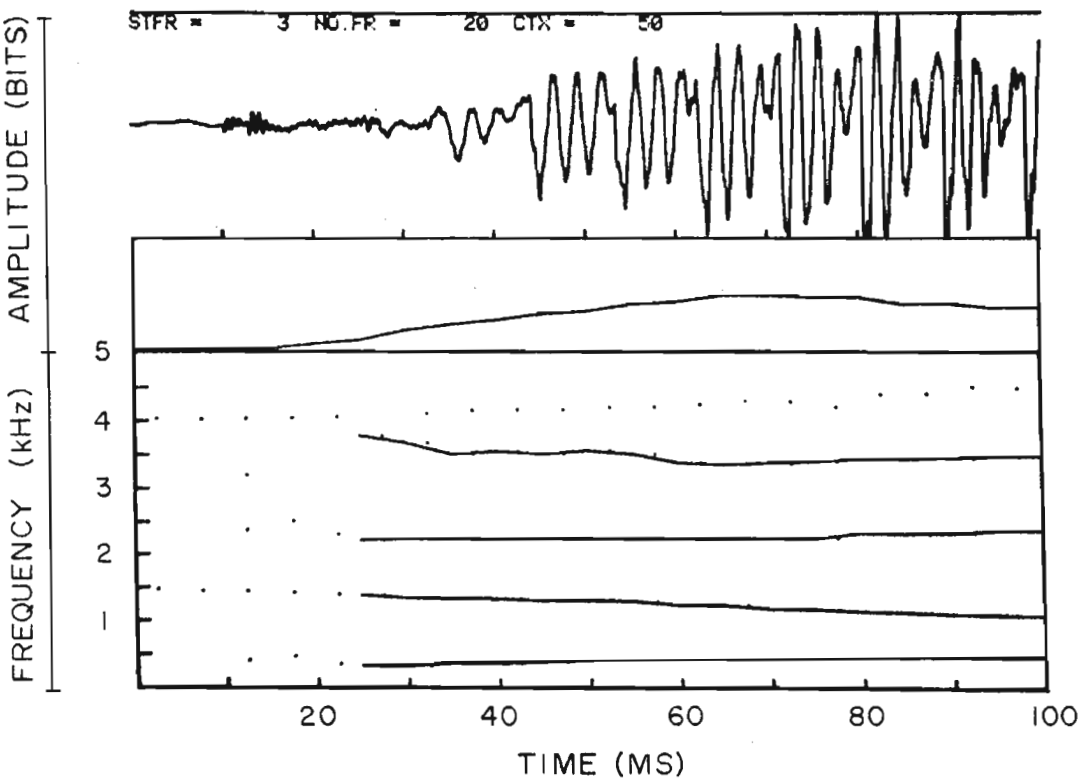


Figure A-3. This figure shows the display of a waveform, the peaks of the spectral sections, the RMS energy and the formant tracks for 20 frames. The coordinates are the same as Figure A-2, except for the presence of the x-axis ticks in 10 ms intervals.

Insert Figures A-4 & A-5 about here

The three-dimensional representation of the running Spectrum for the first eight frames of the waveform is usually examined next. Figure A-4 shows a display of linear prediction smoothed spectra from command TD (three-dimensional display). Figure A-5 also uses TD, but produces spectral sections more like the auditory filters described by Patterson, 1974, using a logrithmic frequency scale.

In summary, SPECTRUM was developed to provide flexible analysis conditions for detailed spectral analysis of speech signals by digital techniques. Its growth has been conditioned by many of the issues raised in this dissertation and other demands in our laboratory. The program has proven to be a user oriented, general purpose research tool which we anticipate will see extensive use in the years to come.

E. Acknowledgements

We are very grateful to Dr. June Shoup-Hummel and the Speech Communication Research Laboratory, Inc. in Los Angeles for permitting us to obtain copies of The Interactive Laboratory System (ILS) programs and command

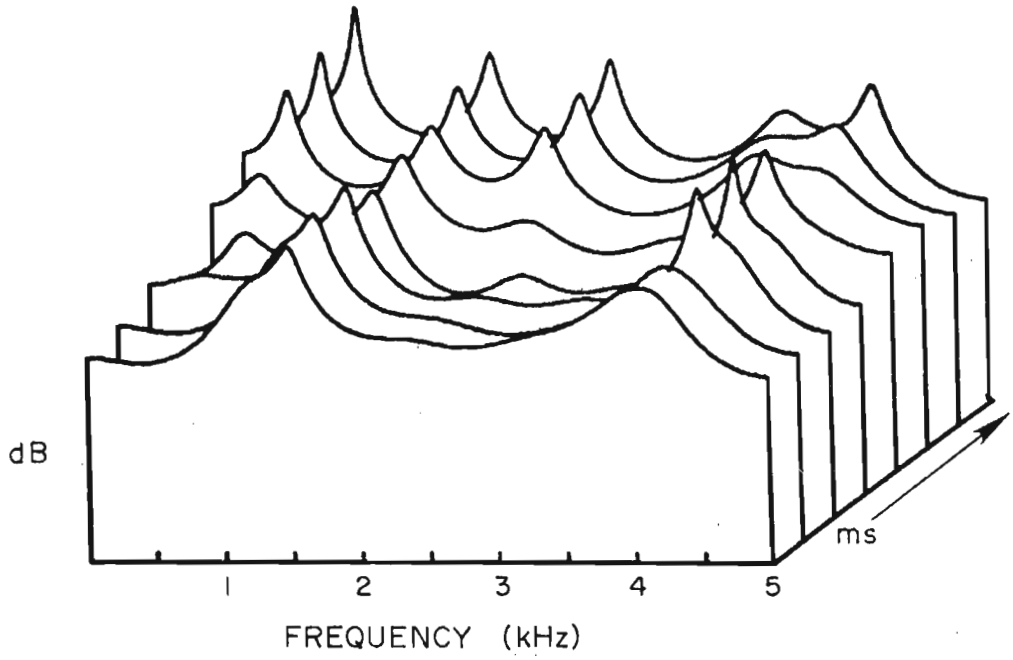


Figure A-4. Three-dimensional display of the smoothed spectra of the inverse filter coefficients at the onset of /go/ are plotted using TD. The x-axis is frequency in 500 Hz ticks. The y-axis is relative dB. The z-axis is in 5 ms intervals which is the frame size for this analysis.

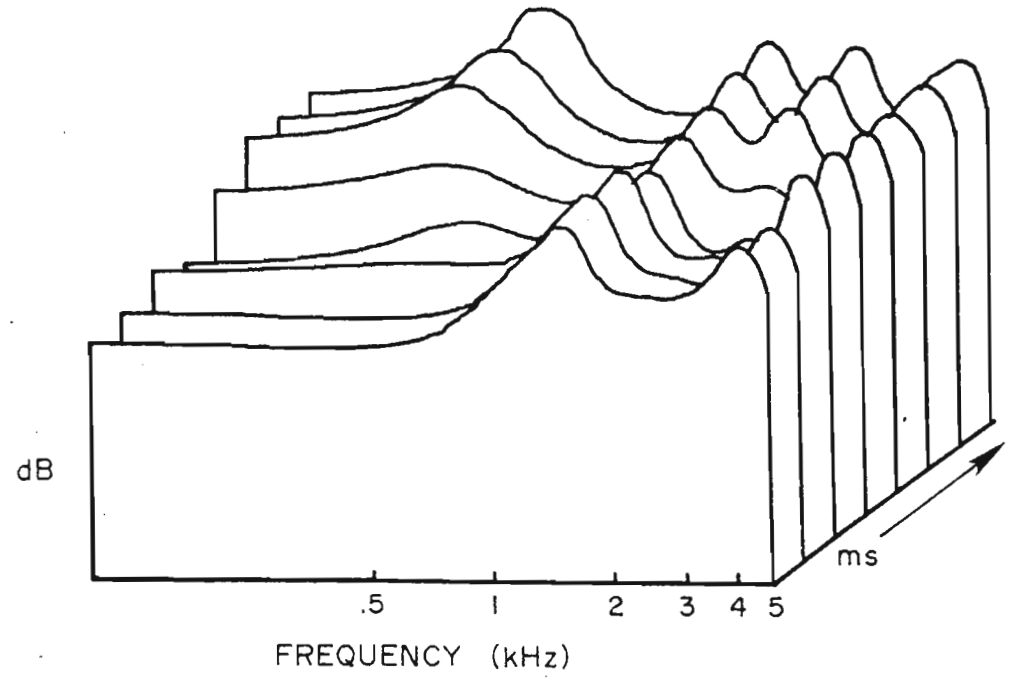


Figure A-5. This display is almost the same as that of Figure A-4, except that the waveform has been analyzed using routines to approximate filtering in the peripheral auditory system.

manuals after which SPECTRUM was closely patterened. We are also grateful to Dr. Steven Davis for providing us with help and assistance with the ILS system during his stay at Haskins Laboratories. Dr. David Broad and Dr. Larry Pfeifer were also helpful in the early stages of this project. The development of SPECTRUM was supported by NIH Grant NS-12179 to Indiana Univerisity.

.F Example of Spectrum command instructions.

SPECTRUM Commands 5/80

****
SMOOTHED SPECTRUM COMMAND SS
****

Function

To calculate the frequency spectrum of the inverse filter and display it for one frame. Can be compared to FFT of frame using XF or XR.

Prompt

STFR, ERASE, LOG, FILTER, DB ->

Command Arguments

STFR = Frame to be displayed
If STFR = Ø, then default and value in COMMON used.
If STFR > Ø, then COMMON value set to STFR.

ERASE = Erase screen before display
If ERASE = Ø, then default. Erase screen before display and save this display on disk for redisplay using SR.
If ERASE = other, keep current display. A total of 3 smoothed spectrum can be displayed simultaneously.

LOG = Log or linear X-axis frequency display
If LOG = Ø, then default and X-axis is linear
If LOG = other, X-axis log scale

FILTER = Auditory Filtering
If FILTER = Ø = default, no filtering applied
If FILTER = other, filtering applied to spectrum as requested.

DB = dB value to add to spectrum to shift spectrum on display frame.
If DB = Ø, normalized gain spectrum displayed
If DB = Ø, then checked that $-5Ø \leq DB \leq +5Ø$, and DB value added to each spectrum value before display.

Global Parameters Used

NBITS = Power of 2 of the FFT. Value of NBITS in GLOBAL COMMON is used to calculate FFT of analysis coefficients, and the resulting smoothed spectrum as stored in the analysis vector (see GP for further information). Default value is 9, or 256 point FFT.

Preparatory Commands

OW A waveform and an analysis file must be open.

AC Analysis coefficients of the inverse filter must be stored in the analysis vector for the frame specified.

Confirmatory Commands

Self-confirming in display

LI can be used to print the smoothed spectrum.

Description

The frequency spectrum of the inverse filter coefficients is a smoothed amplitude-versus-frequency display of the model of the resonances of the vocal tract with the fundamental frequency removed. The SS command calculates the spectrum as log magnitude of the Fourier Transform in dB. The calculated spectrum is stored in the analysis vector in dB and is displayed on the VT-11 using either a linear or log frequency display. The smooth spectrum is always gain normalized by adding the constant GAIN to the spectral values [see AC] and is further shifted by the constant DB when specified by the user.

The smoothed spectrum may be filtered before display by the subroutine OCT13. OCT13 calculates the output of running averaging filters similar to those proposed for the auditory system. (see FILTER) The prompt for OCT13 is:

GROS ANA., BW CON. (REAL), SKIRTS(DB/OCTAVE) ->

Entering 'New Line' specifies the default values for the filter which are for 1/3 octave filters, according to the ANSI S1.11-1966 standards, using the GROS ANALYSIS option. The GROS ANALYSIS option calculates filter outputs less frequently as the bandwidths of the filters increase. Setting the BANDWIDTH CONSTANT to .129 and the SKIRTS to 75 (DB/OCTAVE) approximate the filters proposed by Patterson (1974).

The SS program first reads the analysis vector for the specified frame (if it does not exist, SS aborts with a message). The analysis vector is checked to see whether the spectrum is already stored by previous use of SS or PP for this frame. If spectrum exists, the program proceeds to the display. If not, the analysis coefficients are read and NBITS is obtained from GLOBAL COMMON. A Fast Fourier Transform of the A coefficients is calculated so that the number of values of magnitude on output is 2**(NBITS -1). The magnitude is converted to dB for display and stored in the analysis vector as integer *512. 256 locations in the analysis vector are alloted to the spectrum, which limits NBITS to 9.

Examples

Fig. SS-1 shows the use of SS to contrast spectral sections of two adjacent frames of /bi/ using linear X-axis display. The commands were:

SS (ret)

STFR, ERASE, LOG, FILTER, DB -> 2 (ret)

SS (ret)

STFR, ERASE, LOG, FILTER, DB -> 3, 1 (ret)

The display of frame 2 was stored on the disk for rapid retrieval by SR because ERASE = ∅.

Fig. SS-2 demonstrates how the smoothed spectral sections produced by SS can be compared to an discrete Fourier Transform of the same frame of /da/ using XF and the log display. The commands were:

XF (ret)

STFR, NPOINT, LOG, FILTER, DB -> 15, 2∅∅, 1, (ret)

XF (ret)

STFR, NPOINT, LOG, FILTER, DB -> 15, 2∅∅, 1, 1, -25 (ret)

GROS ANA., BW CON.(REAL), SKIRTS(DB/OCTAVE) -> , .13, 75 (ret)

SS (ret)

STFR, ERASE, LOG, FILTER, DB -> 15, 1, 1, 1, -4∅ (ret)

GROS ANA., BW CON.(REAL), SKIRTS(DB/OCTAVE) ->, .13, 75 (ret)

The first call to XF produced a 2∅∅ point FFT display on a log frequency axis. The second call to XF calculated the same FFT, but filtered it with Patterson type filters before the log display. It was offset by -25 dB for display purposes. The last call to SS showed the smoothed spectrum after it was filtered with Patterson filters, and offset by -4∅ dB.

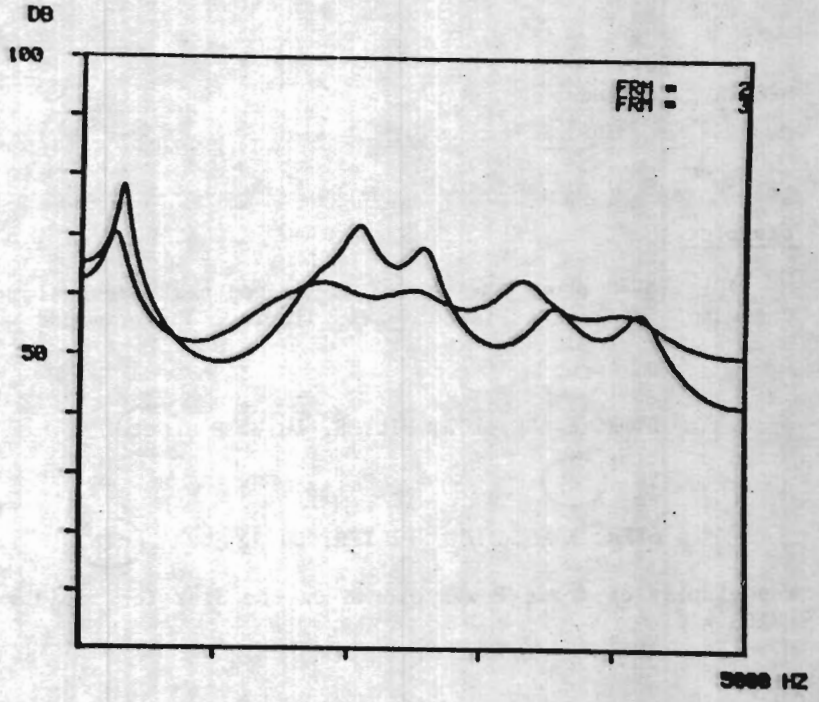


Fig. SS-1. Display showing two spectral sections produced by two calls to command SS.

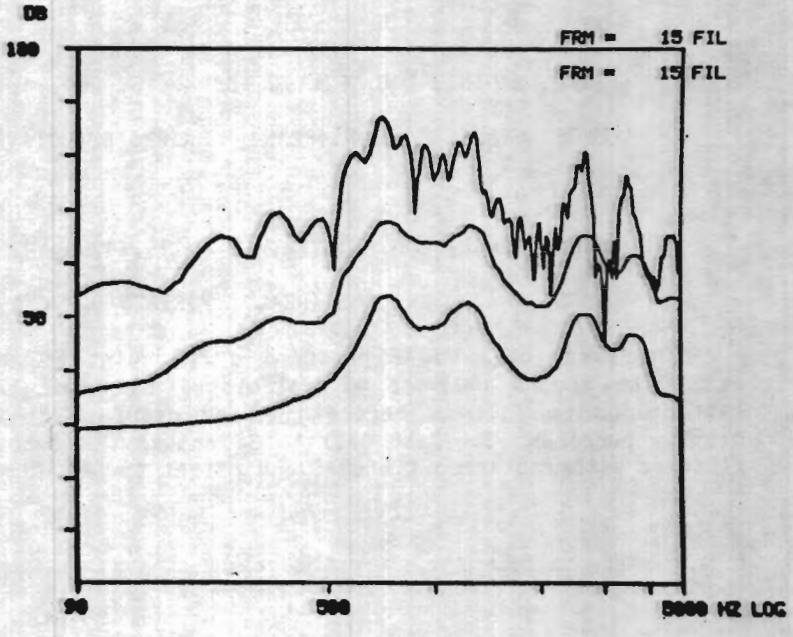


Fig. SS-2. Comparison of spectral sections produced by SS command and discrete Fourier Transform display produced by command XF when auditory filtering is applied.

APPENDIX B. TABLE OF FORMANT TRANSITION PARAMETERS

This appendix contains the average values of the formant transition parameters measured in Experiment 1. All values are averaged over five measurements. The durational measurements are presented in milliseconds, and are converted from the original measurements in frames at 5 ms intervals. All formant transitions were approximated as single straight line segments, except the F3 for /do/. The average values for the 3 segments for F3 of /do/ are presented at the end of the table. Column 1 presents the time of formant onset relative to the burst. The relative onset time for F1 signifies the onset of voicing for the CV, and therefore measures VOT as defined by Lisker and Abramson (1964). Column 5 represents the average value of the frequency of the formant transition 95 ms from the burst, whether or not the steady-state value has been attained.

Table B1. Formant transition parameters.

		Time of onset relative to burst ms	Onset freq. freq. Hz	Trans. duration ms	Vowel steady-state freq. Hz	Freq. 95 ms from burst Hz
/bi/	F1	5	310	14	288	279
	F2	5	1816	32	2123	2300
	F3	5	2462	30	2625	2765
/di/	F1	12	298	9	278	275
	F2	12	1920	41	2137	2277
	F3	12	2587	25	2618	2768
/gi/	F1	17	275	7	263	291
	F2	22	2322	12	2226	2270
	F3	17	2835	21	2795	2782
/bI/	F1	3	359	9	356	363
	F2	3	1745	24	1860	1951
	F3	5	2424	22	2531	2551
/dI/	F1	14	313	17	340	372
	F2	14	1881	17	1916	1966
	F3	14	2587	8	2582	2592
/gI/	F1	23	284	25	322	374
	F2	25	2182	33	2084	2053
	F3	23	2679	32	2538	2538
/be/	F1	4	371	27	459	475
	F2	4	1663	25	1804	1913
	F3	7	2402	41	2531	2582
/de/	F1	10	340	20	411	476
	F2	10	1842	9	1847	1904
	F3	10	2556	17	2582	2558
/ge/	F1	18	340	19	385	450
	F2	21	2174	34	2058	2039
	F3	20	2580	43	2538	2541

Table B1, continued.

		Time of onset relative to burst ms	Onset freq. freq. Hz	Trans. duration ms	Vowel steady-state freq. Hz	Freq. 95 ms from burst Hz
/bɛ/	F1	3	344	24	460	529
	F2	3	1592	39	1736	1758
	F3	5	2302	32	2478	2514
/dɛ/	F1	11	341	35	450	501
	F2	11	1831	5	1815	1807
	F3	11	2562	27	2466	2495
/gɛ/	F1	15	301	43	427	485
	F2	16	2180	63	1878	1866
	F3	21	2566	38	2470	2464
/bae/	F1	2	433	30	615	727
	F2	2	1526	14	1571	1631
	F3	2	2156	34	2387	2394
/dae/	F1	10	370	47	621	652
	F2	10	1786	55	1696	1697
	F3	10	2548	38	2481	2487
/gae/	F1	18	368	63	606	630
	F2	19	2112	71	1771	1756
	F3	27	2416	49	2456	2441
/ba/	F1	2	395	24	646	702
	F2	2	1069	21	1069	1083
	F3	3	2425	38	2590	2646
/da/	F1	9	392	38	669	710
	F2	9	1660	65	1166	1154
	F3	9	2628	30	2523	2522
/ga/	F1	20	384	46	645	660
	F2	20	1733	69	1194	1201
	F3	22	2367	35	2512	2518

Table B1, continued.

		Time of onset relative to burst ms	Onset freq. freq. Hz	Trans. duration ms	Vowel steady-state freq. Hz	Freq. 95 ms from burst Hz
/bo/	F1	1	378	12	440	465
	F2	1	1127	18	1079	993
	F3	2	2379	24	2434	2483
/do/	F1	11	350	29	441	491
	F2	11	1635	79	1190	1179
	F3	11	2549	74	2408	2419
/go/	F1	26	322	42	447	459
	F2	26	1474	67	1134	1142
	F3	26	2274	19	2265	2380
/bu/	F1	4	327	5	354	341
	F2	4	1064	22	1012	983
	F3	4	2300	15	2287	2358
/du/	F1	11	301	17	327	344
	F2	11	1701	34	1536	1340
	F3	11	2430	27	2257	2252
/gu/	F1	25	334	4	328	351
	F2	25	1275	11	1234	1101
	F3	27	2203	6	2154	2230

Three-segment F3 transition for /do/.

		Onset A	Point B	Point C	SS D
F3	Hz	2549	2479	2245	2408
Length	ms		21 ms	10 ms	43 ms

APPENDIX C. FEATURE DEFINITION SHEET FOR RUNNING SPECTRA FROM EXPERIMENT 2

The feature definitions listed below were presented to the judges in Experiment 2 and were available for inspection during the entire experimental session.

A 'spectral section' or 'frame' is a frequency-by-amplitude display of energy averaged over a short time interval.

Feature 1) Tilt of the spectrum at the burst onset below 3500 Hz. Tilt is estimated by visually fitting a straight line to the first frame between 0 and 3500 Hz. Tilt feature categories:

R = rising

F = flat or falling

Feature 2) Late onset of low frequency energy. Late onset is defined as the occurrence of high amplitude, low frequency peaks starting in the fourth frame or later. Late onset feature categories:

L = late onset

N = no late onset

Feature 3) Mid-frequency peaks extending over time. This feature is determined by the existence of a single, prominent peak between 1000 and 3500 Hz occurring for three or more frames, but not necessarily consecutive frames. Mid-frequency features:

Y = yes, peaks exist

N = no such peaks

Now, assign consonant b, d or g according to the feature responses you have made as shown in the following table. An entry of '?' means that either feature category may occur for that stop.

Tilt	Late Onset	Mid-freq. Peaks	Assigned Consonant
F	N	N	b
R	?	N	d
?	L*	Y	g

The '*' by the feature L indicates that in ambiguous cases, the presence of L is sufficient to assign the stop g.

Bibliography

Bailey, P. J., Summerfield, Q. and Dorman, M. On the identification of sine-wave analogues of certain speech sounds. Haskins Laboratories Status Report on Speech Research, SR-51/52, 1977, 1-25.

Bhat, D. N. S. A general study of palatalization. In J. H. Greenberg (Ed.), Universals of Human Language, Volume 2. Stanford: Stanford University Press, 1978, 47-92.

Bladon, R. A. W. and Lindblom, B. Auditory modeling of vowels. In J. J. Wolf and D. H. Klatt (Eds.), Speech Communication Papers Presented at The 97th Meeting of the Acoustical Society of America. New York: Acoustical Society of America, 1979, 1-4.

Blumstein, S. E. and Stevens, K. N. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. Journal of the Acoustical Society of America, 1979, 66, 1001-1017.

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

Carlson, R. and Granstrom, B. Model predictions of vowel dissimilarity. Quarterly Progress and Status Report, STL-QPRS 3-4, Stockholm: Speech Transmission Laboratory, 84-104.

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

Cole, R. A. and Scott, B. The phantom in the phoneme: Invariant cues for stop consonants. Perception & Psychophysics, 1974, 15, 101-107. (a)

Cole, R. A. and Scott, B. Toward a theory of speech perception. Psychological Review, 1974, 81, 348-374. (b)

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

Cullinan, W. L. and Tekieli, M. E. Perception of vowel features in temporally-segmented noise portions of stop-consonant CV syllables. Journal of Speech & Hearing Research, 1979, 22, 122-131.

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

Delgutte, B. Representations of speech-like sounds in the discharge patterns of auditory-nerve fibers. Journal of the Acoustical Society of America, 1980, 843-857.

Dorman, M. F., Studdert-Kennedy, M. and Raphael, L. J. Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. Perception & Psychophysics, 1977, 22, 109-122.

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

Fant, G. Analysis and synthesis of speech processes. In B. Malmberg (Ed.), Manual of Phonetics. Amsterdam: North-Holland, 1968, 173-277.

Fant, G. Stops in CV-syllables. In G. Fant Speech Sounds and Features. Cambridge, MA: MIT, 1973, 110-139.

Fischer-Jørgensen, E. Acoustic analysis of stop consonants. Miscellanea Phonetica, 1954, 2, 42-49.

Fischer-Jørgensen, E. Tape cutting experiments with Danish stop consonants in initial position. Annual Report VII. Copenhagen, Denmark: Institute of Phonetics, University of Copenhagen, 1972, 104-175.

Flanagan, J. L. Parametric coding of speech spectra. Journal of the Acoustical Society of America, 1980, 68, 412-419.

Flanagan, J. L. and Christensen, S. W. Computer studies on parametric coding of speech spectra, Journal of the Acoustical Society of America, 1980, 68, 420-430.

Forshee, J. C. Speech perception laboratory: Current computer resources. RESEARCH ON SPEECH PERCEPTION Progress Report No. 5, Indiana University, 1979, 449-474.

Fowler, C. Coarticulation and theories of extrinsic timing. Journal of Phonetics, 1980, 113-133.

Halle, M., Hughes, G. W. and Radley, J. P. A. Acoustic properties of stop consonants. Journal of the Acoustical Society of America, 1957, 29, 107-116.

Harris, K. S., Hoffman, H. S., Liberman, A. M., Delattre, P. C. and Cooper, F. S. Effect of third formant transitions on the perception of the voiced stop consonants. Journal of the Acoustical Society of America, 1958, 30, 122-126.

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

Jakobson, R., Fant, C. G. M. and Halle, M. Preliminaries to speech analysis. Cambridge, MA: M.I.T., 1952.

Joos, M. Acoustic phonetics. Language, 1948, Supplement 48, 1-137.

Just, M., Suslick, R. L., Michaels, S. and Shockey, L. Acoustic cues and psychological processes in the perception of natural stop consonants. Perception & Psychophysics, 1978, 24, 327-336.

Kewley-Port, D. KLTEXC: Executive Program to implement the KLATT software speech synthesizer. RESEARCH ON SPEECH PERCEPTION Progress Report No. 4, Indiana University, 1978, 235-246.

Kewley-Port, D. Continuous spectral change as acoustic cues to place of articulation. RESEARCH ON SPEECH PERCEPTION Progress Report No. 5, Indiana University, 1979, 327-346. (a)

Kewley-Port, D. Spectral continuity of burst and formant transitions as cues to place of articulation in stop consonants. In J. J. Wolf and D. H. Klatt (Eds.), Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America, New York: Acoustical Society of America, 1979, 175-178. (b)

Kiang, N. Y. S. Processing of speech by the auditory nervous system, Journal of the Acoustical Society of America, 1980, 68, 830-835.

Kiang, N. Y. S., Eddington, D. K. and Delgutte, B. Fundamental considerations in designing auditory implants. Acta Otolaryngology, 1979, 87, 204-218.

Klatt, D. H. Voice onset time, frication, and aspiration in word-initial consonant clusters. Journal of Speech and Hearing Research, 1975, 18, 686-706.

Klatt, D. H. A digital filter bank for spectral matching. In C. Teacher (Ed.), Conference Record of the 1976 IEEE International Conference on Acoustics Speech and Signal Processing. Philadelphia, PA, IEEE Catalog No. 76CH1067-8 ASSP, 1976, 537-540.

Klatt, D. H. Speech Perception: A Model of Acoustic-Phonetic Analysis and Lexical Access. Journal of Phoentics, 1979, 7, 279-312.

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

Klatt, D. H. Lexical representations and processing strategies during speech production and perception. Psychological Review, 1980b (In Press).

Klein, W., Plomp, R. and Pols, L. C. W. Vowel spectra, vowel spaces, and vowel identification. Journal of the Acoustical Society of America, 1970, 48, 999-1009.

LaRiveriere, C., Winitz, H. and Herriman, E. Vocalic transitions in the perception of voiceless initial stops. Journal of the Acoustical Society of America, 1975, 57, 470-475.

Lehiste, I. and Peterson G. E. Transitions, glides and diphthongs. Journal of the Acoustical Society of America, 1961, 33, 268-277.

Liberman, A. M. The grammars of speech and language. Cognitive Psychology, 1970, 1 (4), 301-323.

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

Liberman, A., Delattre, P. C. and Cooper, F. Some cues for the distinction between voiced and voiceless stops in initial position. Language and Speech, 1958, 1, 153-167.

Liberman, A. M., Delattre, P. C., Cooper, F. S. and Gerstman, L. J. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychological Monographs, 1954, 68 (8, Whole No. 379), 1-13.

Liberman, A. M., Mattingly, I. G. and Turvey, M. T. Language codes and memory codes. In A. W. Melton and E. Martin (Eds.), Coding Processes in Human Memory. New York: V. H. Winston & Sons, 1972, 307-334.

Liberman, A. M. and Pisoni, D. B. Evidence for a special speech perception subsystem in the human. In T. H. Bullock (Ed.), Recognition of Complex Acoustic Signals. Berlin: Dahlem Konferenzen, 1977, 59-76.

Liberman, A. M. and Studdert-Kennedy, M. Phonetic Perception. In R. Held, H. Leibowitz and H. L. Teuber (Eds.), Handbook of Sensory Physiology: Perception. New York: Springer-Verlag, 1978, 143-179.

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

Lindblom, B. E. F. and Studdert-Kennedy, M. On the role of formant transitions in vowel recognition, Journal of the Acoustical Society of America, 1967, 42, 830-843.

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

Markel, J. D. FFT pruning. IEEE Transactions, 1971, AV-19, 305-311.

Markel, J. D. and Gray, A. H. Linear prediction of speech. New York: Springer-Verlag, 1976.

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

Miller, J. L. The perception of voicing and place of articulation in initial stop consonants: Evidence for the nonindependence of feature processing. Journal of Speech and Hearing Research, 1977, 20, 519-528.

Monsen, R. On the measurability of vowel formants: or the recoverability of vocal-tract resonances in the sound wave, 1981 (in preparation).

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K. and Bent, D. H. SPSS: Statistical Package for the Social Sciences, New York: McGraw-Hill, 1975.

Oden, G. C. and Massaro, D. W. Integration of featural information in speech perception, Psychological Review, 1978, 85, 179-191.

Ohde, R. N. and Sharf, D. J. Order effect of acoustic segments of VC and CV syllables on stop and vowel identification. Journal of Speech and Hearing Research, 1977, 20, 543-554.

Öhman, S. E. G. Coarticulation in VCV utterances: Spectrographic measurements. Journal of the Acoustical Society of America, 1966, 39 (1), 151-168.

Patterson, R. D. Auditory filter shape. Journal of the Acoustical Society of America, 1974, 55, 802-809.

Patterson, J. H. and Green, D. M. Discriminability of transient signals having identical energy spectra. Journal of the Acoustical Society of America, 1970, 48, 894-905.

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

Port, R. F. On the structure of the phonetic space with special reference to speech timing. 1980 (Manuscript submitted for publication.)

Potter, R. K., Kopp, G. A. and Green, H. C. Visible Speech, 1947, New York.

Rabiner, L. R. and Gold, B. Theory and Application of Digital Signal Processing. 1975, Englewood Cliffs, NJ: Prentice Hall.

Sachs, M. B. and Young, E. D. Effects of nonlinearities on speech encoding in the auditory nerve. Journal of the Acoustical Society of America, 1980, 68, 858-875.

Sachs, R. M. and Grant, K. W. Stimulus correlates in the perception of voice onset time (VOT): II. Discrimination of speech with high and low stimulus uncertainty. Journal of The Acoustical Society of America, 1976, 60(1), S91(A).

Sawusch, J. R. and Pisoni, D. B. On the identification of place and voicing features in synthetic stop consonants. Journal of Phonetics, 1974, 2, 181-194.

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

Schouten, M. E. H. and Pols, L. C. W. CV- and VC-transitions: a spectral study of coarticulation - Part II. Journal of Phonetics, 1979, 7, 205-224.

Schroeder, M. R, Atal, B. S. and Hall, J. L. Optimizing digital speech coders by exploiting masking properties of the human ear. Journal of the Acoustical Society of America, 1979, 66, 1647-1652.

Searle, C. L., Jacobson, J. Z. and Rayment, S. G. Stop consonant discrimination based on human audition. Journal of the Acoustical Society of America, 1979, 65, 799-809.

Searle, C. L., Jacobson, J. Z. and Kimberly B. P. Speech as patterns in the 3-space of time and frequency. In R. A. Cole (Ed.), Perception and Production of Fluent Speech, Hillsdale, NJ: Erlbaum, 1980, 73-102.

Sever, J. C. and Small A. M. Binaural critical masking bands. Journal of the Acoustical Society of America, 1979, 66, 1343-1350.

Sharf, B. Critical bands. In J. V. Tobias (Ed.), Foundations of Modern Auditory Theory, Vol. 1. New York: Academic Press, 1970, 157-202.

Stevens, K. N. Acoustic correlates of certain consonantal features. Paper presented at Conference on Speech Communication and Processing, MIT, Cambridge, MA: November 6-8, 1967.

Stevens, K. N. Segments, features and analysis by synthesis. In J. F. Kavanagh & I. G. Mattingly (Eds.), Language by ear and by eye. Cambridge, MA: M.I.T., 1972.

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

Stevens, K. N. Acoustic correlates of some phonetic categories. Journal of the Acoustical Society of America, 1980, 68, 836-842.

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

Stevens, K. N. and Blumstein, S. E. The search for invariant acoustic correlates of phonetic features. In P. D. Eimas and J. Miller (Eds.), Perspectives on the Study of Speech, 1980 (In Press).

Stevens, K. N. and House, A. S. Studies of formant transitions using a vocal tract analog. Journal of the Acoustical Society of America, 1956, 28, 578-585.

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

Strange, W., Verbrugge, R., Shankweiler, D. and Edman, T. Consonant environment specifies vowel identity. Journal of the Acoustical Society of America, 1976, 60, 213-224.

Studdert-Kennedy, M. Universals in phonetic structure and their role in linguistic communication. In T. H. Bullock (Ed.), Recognition of Complex Acoustic Signals. 1977. Berlin: Dahlem Konferenzen, 37-48.

Studdert-Kennedy, M. Perceiving phonetic segments. In T. F. Myers, J. Lover, and J. Anderson (Eds.), The Cognitive Representation of Speech, Amsterdam: North-Holland, 1980 (In Press).

Tekieli, M. E. and Cullinan, W. L. The perception of temporally segmented vowels and consonant-vowel syllables. Journal of Speech & Hearing Research, 1979, 22, 103-121.

Winitz, H., Scheib, M. E. and Reeds, J. A. Identification of stops and vowels for the burst portion of /p,t,k/ isolated from conversational speech. Journal of the Acoustical Society of America, 1972, 51, 1309-1317.

Young, E. D. and Sachs, M. B. Representations of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. Journal of the Acoustical Society of America, 1979, 66, 1381-1403.

Zhukov, S. Ya., Zhukova, M. G. and Chistovich, L. A. Some new concepts in the auditory analysis of acoustic flow. Soviet Physical Acoustics, 1974, 20, 237-240 [Akus. Zh. 20, 386-392 (1974)]

Zue, V. W. Acoustic characteristics of stop consonants: A controlled study. Technical Report No. 523, Lincoln Laboratory, M.I.T., May 1976.

Zwicker, E. and Feldtkeller, R. Das Ohr als Nachrich tenempfanger, Stuttgart: Hirzel, 1967.

Zwicker, E., Terhardt, E. and Paulus, E. Automatic speech recognition using psychoacoustic models. Journal of the Acoustical Society of America, 1979, 65, 487-498.