University of Alberta

The Perception of Spectrally and Temporally Distorted Prevocalic Stop Consonants

by

Michael Kiefte ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in

Experimental Phonetics

Department of Linguistics

Edmonton, Alberta
Fall 2000

0-612-59611-7

University of Alberta

Library Release Form

Name of Author: Michael Kiefte

Title of Thesis: The Perception of Spectrally and Temporally Distorted Prevocalic Stop Consonants

Degree: Doctor of Philosophy

Year this Degree Granted: 2000

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material from whatever without the author's prior written permission.

133 Patrick Street
St. John's, Newfoundland
A1E 2S9 Canada

July 24, 2000

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled The Perception of Spectrally and Temporally Distorted Prevocalic Stop Consonants submitted by Michael Kiefte in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Experimental Phonetics.

Dr. Terrance Nearey,
Supervisor

Dr. Joseph Pater,
Committee Member

Dr. Michael Dawson,
Committee Member

ROEL SMITS

Dr. Roel Smits,
Committee Member

Dr. John Kingston,
External Committee Member

John T. Hogan

Dr. John Hogan,
Committee Member

24 July 2000

Abstract

Two sets of experiments are designed to evaluate several proposed acoustic features as potential correlates to place of articulation in prevocalic stop consonants. In both cases, naturally produced stimuli are distorted so that one set of cues is removed in order to evaluate its perceptual importance.

In the first set of experiments, temporal information in short, gated stops is distorted by altering the waveform envelope while preserving the long-term power spectrum. It is assumed that if temporally dynamic information is phonetically relevant in the perception of such stimuli, then this signal processing should have an adverse effect on correct identification performance. It is found that the release burst can be treated as a static spectral unit up to a duration of approximately 20 ms unless the voice onset time (VOT) is shorter than this. It is, however, concluded that short, gated stop bursts cannot be treated as an indivisible unit irrespective of VOT.

The second set of experiments addresses the importance of gross *vs.* detailed spectral features in the perception of stop-consonant place of articulation. This is done by presenting listeners with noise-vocoded stimuli in which the spectral resolution of the processed tokens is severely reduced. While it can be assumed that such a transformation will effectively eliminate detailed spectral features, such as burst peak and formant frequencies, it is shown that this is not necessarily the case. Several simulations demonstrate that much detailed cue information is potentially recoverable from noise-vocoded stimuli even when the noise-vocoding bandwidth is extremely broad. Nevertheless, it is shown that gross spectral features, such as spectral tilt and compactness, are better able to predict listeners' identifications of noise-vocoded stimuli. This result is subsequently shown to be true for both the burst and the vocalic formant transitions that

follow: *i.e.*, correct identification rates for burtless stimuli remain significantly above chance levels even at very wide noise-vocoding bandwidths. While this suggests that listeners attend primarily to gross spectral shape cues which are assumed to be particularly robust in noise-vocoded speech, it is also shown that detailed spectral features are better able to model listeners' responses to undistorted speech.

Acknowledgements

This work could not have been possible without the tremendous support of two people who have encouraged me through the last seven years. Firstly, my wife Kerry-Joy Kiefte, has worked as much on this dissertation as I have; without her support as a partner, I would not have had the endurance to complete my graduate studies, either emotionally or physically. Although she does not receive any of the credit for this work directly, this thesis is as much a part of her accomplishments as it is of mine.

I have also been fortunate enough to have worked with the best supervisor I could have asked for. Terry Nearey's knowledge of Phonetics and Statistics has been a tremendous resource over the years and will continue to have an important influence on my research for many years to come. Very importantly, he has allowed me to explore many of my own ideas in methodology, analysis, and phonetic theory, while subtly steering my approach in an informed and experienced manner. Without his influence, this work would have taken on an entirely different form and would therefore have suffered enormously. I can only hope that his role as a mentor will not end with the completion of this dissertation and that I can look forward to continued collaboration with him.

I consider myself extremely lucky to have been able to rely on these two people.

I would also like to thank two other people who have additionally played an important role in my life—albeit much more recently and for much less time. I would like to thank David Kiefte for waiting patiently while I finished something called a "thesis" before asking to play outside or ride our bikes. In addition, Andrew Kiefte proofed many earlier versions of this thesis—mostly upside-down—and tested the quality and strength of the paper I was using at the time.

Table of Contents

List of Tables

List of Figures

Chapter 1

The production and perception of oral stop consonants

1.1 Introduction

The perception and identification of the oral stop consonants [p], [t], [k], [b], [d], and [g] have been used as a test bed for competing theories of speech perception since the early 1950s. Examination of the acoustic properties of these sounds has revealed highly context dependent relationships and experiments in stop consonant perception by human listeners have shown relatively complex response patterns. Because they possess acoustic and articulatory characteristics found in most other speech sounds—*e.g.*, the noisy frication energy found in consonants such as [s] and [f], and the dynamic vocal tract resonances of glides and diphthongs (Sec. 1.2.1)—a better understanding of stop consonants would benefit research in all areas of speech perception.

These six speech sounds are commonly subclassified along two distinct features: voicing and place. The present dissertation focuses on the perception of stop-consonant *place of articulation* with the assumption that there is little or no interaction between this property and *voicing*, either in production or perception. Although it is known that the relative onset of voicing is a *secondary* cue to place-of-articulation identification, it has been shown that this effect is linearly factorable from other place cues (Oden and Massaro, 1978; Benkí, 1998). Therefore, this thesis will only address listeners' ability to discriminate between the homorganic pairs [p,b], [t,d], and [k,g]. Furthermore, this research will concentrate primarily on the voiced stops [b], [d], and [g] for reasons that will be described in Sec. 1.2.1.

There exists strong disagreement regarding the exact nature of the relevant features used by humans to discriminate between different places of articulation in speech perception. Smits, ten Bosch, and Collier (1996a) have grouped the acoustic cues proposed by currently existing hypotheses into two main categories:

gross cues spectral features that are broadly localized in frequency or time. Examples of such cues include overall spectral shape or its relative change over time.

detailed cues features that are narrowly localized in frequency or time, such as the center frequency of prominent spectral peaks and their dynamic transitions.

The appeal of gross spectral shape cues in automatic speech recognition (ASR) stems from the ease with which they can be extracted from running speech. It is known however, that applications using such features are not robust in the context of spectral distortion to which human listeners can easily adapt (Allen, 1994). Additionally, experiments in automatic speech recognition bear no direct relationship to human perception (Nearey, 1992a).

The primary difference between detailed and gross spectral cues is the minimum spectral resolution required to represent them—*i.e.*, the precision with which spectral measurements must be made in order to effectively capture the relevant frequency information. Smits *et al.* (1996a) suggest that detailed cues are sufficiently preserved using a spectral resolution less than 500 Hz. Gross spectral measures, such as the overall spectral slope (Stevens and Blumstein, 1978; Blumstein and Stevens, 1979, 1980; Lahiri, Gewirth, and Blumstein, 1984), should be fully representable with a frequency resolution much broader than this. Based on such a theory of speech perception, place identification in the context of *reduced frequency selectivity*—*e.g.*, a reduction in the spectral resolution of the human auditory system—is expected to be quite robust. Conversely, a detailed cue theory should predict that perception will degrade quite quickly under such conditions.

The effects of reduced frequency selectivity can be observed in hearing-impaired listeners or in recipients of *cochlear implants*. For example, in a perceptual experiment in which both detailed (formant transitions) and gross (spectral tilt) cues were manipulated independently, Lindholm *et al.* (1988) found that place-of-articulation perception by hearing-impaired individuals was closely correlated with gross-spectral-shape properties, whereas responses from normal-hearing subjects corresponded better to detailed acoustic cues. Place identification by hearing-impaired subjects was also much less consistent suggesting that spectral-shape properties are nevertheless not particularly useful in the context of reduced frequency selectivity characteristic of this kind of sensorineural hearing loss. Several researchers have *emulated* such deleterious effects to study the perception of speech by these individuals (*e.g.*, Summers, 1991; Boothroyd *et al.*, 1996; Shannon *et al.*, 1995). Similarly, the purpose of the present dissertation is to compare *gross vs. detailed* theories of stop-consonant perception by simulating the effects of reduced frequency selectivity in stimuli that are presented to normal-hearing listeners for place-of-articulation identification. In principle, theories based on gross spectral cues should predict highly robust consonant identification in spectrally distorted speech, while detailed-cue theories should predict a substantial degradation in performance.

In practice, an important step in evaluating the relative importance of such cues is the statistical modeling of subjects' responses using features[1] proposed by specific theories (*e.g.*, Krull, 1990; Nearey, 1995; Smits, ten Bosch, and Collier, 1996b). The approach taken in this dissertation will be to treat human speech perception as *statistical pattern recognition* (Nearey, 1997). This enables us to evaluate acoustic cues using methods found elsewhere in the statistical classification literature. As such, the problem of speech perception can be divided into two separate issues: the classification algorithm and the features used by this algorithm to discriminate categories (Ripley, 1996). In this thesis, we concentrate on the actual features themselves.

The distortion of potential acoustic cues in stop consonants should help evaluate

[1] The term "feature" is used here purely in the context of statistical pattern recognition—*i.e.*, a continuous dimension of the pattern space. This is not to be confused with the traditional linguistic meaning of "distinctive feature". When the latter meaning is intended, it will be explicitly denoted as such.

their relative importance and will hopefully determine the sufficiency of either gross or detailed acoustic cues. There are two possible outcomes from this kind of research:

1. *A particular signal manipulation has no effect on listeners performance.* This indicates that the distortion does not alter *phonetically relevant* features and that any such features are restricted to the subspace of signal properties that can be shown to remain unaffected by the spectral processing. This delimits the set of possible important spectral cues. This does not necessarily imply that there is no *perceptible* difference between the unprocessed and distorted speech; rather, it means that the signal manipulation does not affect listeners' ability to classify purely phonetic properties, such as stop place of articulation.

2. *A particular signal manipulation has an adverse effect on listeners' performance.* This indicates that at least some part of an important phonetic feature is affected by the signal manipulation. In this situation, it is important to model listeners' responses empirically using explicit pattern-recognition models based on theories of human perception.

Parametric modeling allows the experimenter to not only estimate the relative importance of a specific set of features, but also to evaluate different theories by comparing their predictive power. This will hopefully lead to firm conclusions regarding the validity of specific gross or detailed cues in consonant perception.

This chapter provides an introduction to past research on the production and perception of prevocalic stop consonants. Section 1.2 gives a brief description of the articulatory events that are involved in production. Section 1.3 provides an overview of *detailed* acoustic cues that have been proposed as important perceptual correlates to place identification with an emphasis on voiced stops. It is found that the complexity of these cues has led several researchers to propose a strong relationship between perception and articulation. One specific theory along these lines is the *Motor Theory* of speech perception, which is presented in Sec. 1.4 and is subsequently contrasted with auditory theories of speech perception. Gross-spectral-cue theories are presented in Sec. 1.5 and a brief overview of explicit comparisons of the relative importance of detailed *vs.* gross spectral features is presented in Sec. 1.6. Finally, the goals of the present thesis are discussed in Sec. 1.7.

1.2 Production and acoustics of stop consonants

This section briefly describes the articulatory events that occur during the production of the oral stop consonants as well as some of the acoustic consequences that can be predicted from these actions based on an *acoustic theory of speech production* (Sec. 1.2.1). Some early *spectrographic* studies of production acoustics are also described (Sec. 1.2.2).

1.2.1 Stop production

The articulatory and acoustic events that make up the production of prevocalic stop consonants can be grouped into five distinct phases: closure, release, frication, aspiration and vocalic formant transitions (Fant, 1973; Dorman, Studdert-Kennedy, and Raphael, 1977). The first of these phases distinguishes this class of speech sounds from all others: closure is characterized acoustically by an initial silent hiatus that results from the cessation of air flow from the lungs through the oral cavity caused by an occlusion created by either the lips or the tongue. The stop consonants in English can be further subdivided into three *places of articulation* according to the location of this occlusion within the oral cavity. To produce the alveolar stops, [d] and [t], either the tongue blade or apex makes contact with the alveolar ridge[2], while for the velars [k] and [g], the tongue dorsum is raised towards the velum, also known as the soft palate. The primary articulators for the bilabials [p] and [b] are the lips. Although it has been shown that the duration of this silent period contributes to the perceptual discrimination of the different places of articulation in medial position—*i.e.*, in vowel + consonant + vowel sequences (VCV) —(Repp, 1984), it is impossible to assess its significance in isolated consonant + vowel (CV) syllables since the silence is unbounded. It will therefore not be addressed here.

During the closure period, intraoral air pressure builds and is finally released when the primary articulator—either the lips or the tongue—is withdrawn. This results in a transient noise or "pop" whose acoustic properties are largely determined by the impulse response of the vocal tract at the exact moment of release. Because this impulse response overlaps with the following, much louder period of noisy frication, this acoustic event is rarely studied separately (Sec. 1.3.1).

Although air is permitted to flow through the oral cavity immediately following release, the primary articulator briefly acts as an acoustic spoiler which generates turbulence in the vicinity of the narrow constriction (Stevens, 1971). The resonant spectral properties of this frication are determined primarily by the size and shape of the oral cavity anterior to the point of constriction (Fant, 1970; Stevens, 1993); therefore, some general predictions can be made regarding the spectral shape of the resulting noise based on an acoustic theory of speech production. Because the front cavity for alveolars is roughly 1–2 cm in length, a spectral prominence is expected somewhere between 4–7 kHz, while for velars, where the length of this cavity ranges from 3–7 cm depending on the following vowel, the spectral prominence is expected to be much lower in frequency (Fant, 1970). Since the size of the front cavity is negligible for bilabials, no dominant spectral peak should occur for either [p] or [b] and the frication phase is either weak or absent for these stops (Fant, 1970, 1973; Zue, 1976; Dorman *et al.*, 1977). Because the release transient and subsequent frication are mostly indistinguishable when viewed on a spectrogram or waveform, they are typically analyzed as a single acoustic unit referred to as the "burst". This convention will be followed throughout the present dissertation.

The acoustic source of the frication phase is located primarily in the vicinity of the constriction itself. However, once the articulator is further retracted, the turbulent noise abates and the vocal tract begins to resonate at characteristic frequencies or *formants* which are seen on a spectrogram as relatively high concentrations of spectral energy within a restricted frequency region. The source of vocal resonance is instead either the low energy noise produced by the flow of air from the lungs through the open glottis, or semi-periodic voicing if the glottis is adducted resulting in vocal phonation. In the case of the English voiced stops [b], [d], and [g], phonation begins either before oral release, which results in a low frequency "voice bar" during the otherwise silent

[2] in languages other than English, it is common for the tongue to make contact with the upper teeth resulting in a *dental* consonant [d̪] or [t̪]

period[3], or sometime during frication. In both cases, the resonant source immediately following frication is the glottal voicing.

In the case of the English voiceless stops [p], [t], and [k], however, the onset of glottal phonation does not occur until much later following the end of frication (Lisker and Abramson, 1964, 1967). The resulting intermediate period of voiceless resonance is referred to as *aspiration*. This phase is often associated with either the preceding burst (*e.g.*, Ohde and Sharf, 1977) or the voiced formants that follow (*e.g.*, Ostreicher and Sharf, 1976), as it has properties in common with each: a noisy source similar to frication (originating instead at the glottis) and vocal tract resonances similar to the voiced formant transitions. The voiced stops in English are generally thought to lack this phase. Because of the inconsistent assignment of aspiration to either the burst or transition phases, research on place perception presented in this review will focus primarily on the voiced stops[4].

Some authors do explicitly label a distinct aspiration phase following the burst in voiced stops, however. Experiments which attempt to evaluate the perceptual significance of such a segment—*i.e.*, the period of time between the nominal end of frication energy and the onset of voicing—find that they provide very little phonetic information (Dorman *et al.*, 1977; Just *et al.*, 1978). In addition, it has been observed that "voiced aspiration" segments, as defined in this manner, are often nearly silent or otherwise do not contain clear formant energy (Just *et al.*, 1978). This dissertation will follow the convention of not treating voiced stop aspiration segments as distinct acoustic units (*cf.*, Pols, 1979, who also studies English stops).

The *formant transitions*—or dynamic changes in resonant frequency over time—reflect shifts in articulatory configuration from the narrow constriction to the following vowel target and are present in both the aspiration and voiced phases following the burst. A typical wideband spectrogram illustrating many of these phases is presented in Fig. 1.1 on the following page.

1.2.2 Stop acoustics and early spectrographic studies

The acoustic consequences of these events were not well understood until the invention of the *sound spectrograph* which was capable of producing images that represented spectral energy in both time and frequency. One of the original uses of these speech *spectrograms* was as an aid for the deaf (Potter, Kopp, and Green, 1947). Although their original goal was not at all successful, Potter *et al.* did illustrate several important acoustic correlates to place of articulation in prevocalic stop consonants that have proven useful in reading spectrograms. Ultimately many of these features have been proposed as important *perceptual* correlates in detailed theories of speech perception (Sec. 1.3).

Of primary importance in their program was the concept of a "hub" which was defined as the position of the second formant F_2 along the frequency axis. Although F_2 was readily apparent for the vowels, which were characterized by very slowly moving

[3] see also Barry (1984) who suggests that the voice bar may have perceptually important spectral properties of its own

[4] In languages other than English, it may be the case that the stop consonants traditionally classified as *voiceless* have such short VOTs that they are very similar to the English voiced stops (Lisker and Abramson, 1964). In such cases, the perception of English voiced stops will be compared to that of the voiceless stops of these languages and will collectively be referred to as "short-lag VOT stops" to avoid confusion. This is notably the case for Dutch which further only has two places of articulation for the fully voiced (or prevoiced) stops: [b] and [d].

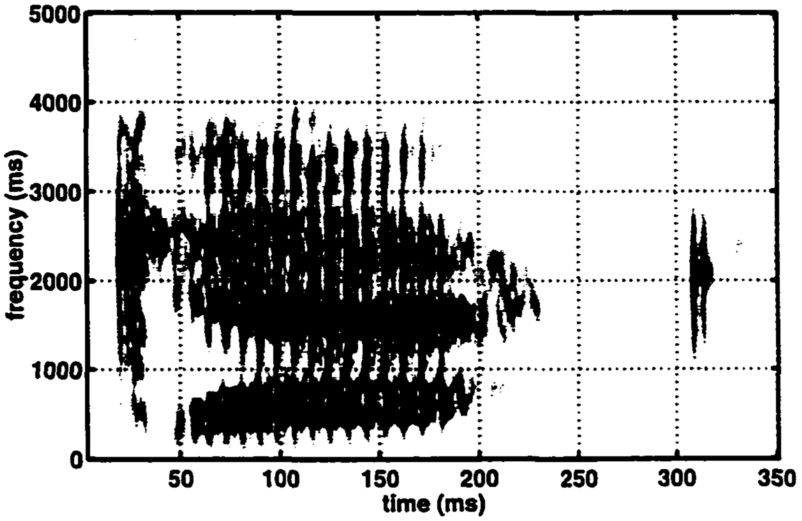


Figure 1.1: Typical wideband spectrogram illustrating many of the stages of prevocalic stop consonant production. This image shows a spectrogram of the nonsense syllable [gæk] as produced by a male speaker using a 6.3 ms Hamming analysis window with an approximate equivalent rectangular bandwidth of 320 Hz.

vocal resonances, the hub was not obvious for prevocalic stops which showed rapid formant transitions. It was claimed that the hub for stops was actually located at the release burst and was therefore obscured; its position had to be inferred by the direction of the second formant transition into the vowel—*i.e.*, a rising or falling F_2 indicated a consonant hub that was respectively lower or higher than the F_2 of the following vowel (Potter *et al.*, 1947, p. 42).

Two important observations were made by Potter *et al.* (1947): that hubs for the three different places of articulation were partly contrastive and that each hub was similar for both voiced and voiceless counterparts. This latter observation was later supported by Fant (1973, p. 118) who demonstrated that, although voiceless stops tended to show slower transitions, formant patterns were very similar *within* places of articulation when tracked backwards through the aspiration segment. By looking for vowel contexts in which F_2 showed little change in frequency during the transition, it was possible to estimate the characteristic hub for each place of articulation. In this manner, it was found that the bilabial hub was located near the F_2 of [u] (Potter *et al.*, 1947, p. 84) and that the hub for alveolars was close to the typical F_2 value for [æ] (p. 90)—*i.e.*, [bu] and [dæ] showed nearly flat F_2 transitions. This can be compared with average values of 1020 Hz and 1720 Hz for the F_2's of [u] and [æ] respectively as observed for adult male speakers (Peterson and Barney, 1952). In contrast, second formant transitions following a velar release were never flat; while the hub for alveolars and bilabials appeared to be fixed, velar F_2 transitions were found to be always falling irrespective of the identity of the following vowel (Potter *et al.*, 1947, p. 97).

Fischer-Jørgensen (1954) found similar patterns in formant transitions for Danish CV syllables. She estimated the hubs for alveolar consonants at 1800 Hz. While this is consistent with other observations regarding the locus for [d] and [t], she also estimated the bilabial locus at 1300 Hz, which is much higher than that found elsewhere [however, Kewley-Port (1982) found even higher values for bilabials before front vowels (Sec. 1.3.2.1)]. For velars, Fischer-Jørgensen found that the second formant transition was actually rising before the front, rounded vowels [y], [ø], and [œ]—none of which occur in English.

Although the observations made by Potter *et al.* (1947) were intended to facilitate the reading of speech spectrograms, the hub concept (later called the "locus") was subsequently proposed by Delattre, Liberman, and Cooper (1955) as an important perceptual correlate to place of articulation identification (see Section 1.3.2.1).

1.3 Detailed acoustic properties

Because detailed-cue theories are described in terms of temporally localized spectral features, it is typically the case that the burst and formant transitions are treated as distinct perceptual properties. Section 1.3.1 describes previous experiments that have been designed to evaluate the effect of release bursts on listeners' ability to discriminate place of articulation, while Sec. 1.3.2 discusses similar experiments involving formant transitions.

Because of the distinction made between burst and formant transitions, such detailed-cue theories must also address the problem of how disparate acoustic features are integrated to form a single perceptual unit. Research that examines the sufficiency, necessity, and relative importance of specific cues can help build models designed to further evaluate their perceptual relevance. Therefore, Sec. 1.3.4 describes experiments that evaluate the relative perceptual importance of these detailed acoustic features.

Additional acoustic properties that have been proposed as possible correlates to place perception are presented in Sec. 1.3.3.

1.3.1 Acoustic properties and perception of the release burst

Although Potter *et al.* (1947) made very little mention of the spectral properties of the burst, they did note that most of the energy for bilabial releases was found towards the lower end of the frequency scale, while velar bursts showed a high concentration of spectral energy near its characteristic hub—*i.e.*, just above the F_2 of the following vowel[5]. Following similar informal investigations, Liberman *et al.* (1952) designed synthetic bursts using a machine called the *Pattern Playback* which was capable of converting highly schematized, hand-painted spectrograms into sound (Cooper, 1950). The perceptual effects of the burst were studied by parametrically manipulating their spectral properties. In this particular experiment, the bursts were painted as 600 Hz wide "teardrop" shapes centered at twelve different frequencies, each of which were presented in all of seven synthetic, two-formant vowel contexts (with F_2 from highest to lowest): [i], [e], [ɛ], [ɑ], [ɔ], [o], and [u]. The resulting stimulus continuum was then presented to listeners who were asked to identify the stimuli as containing either [p], [t], or [k].

The modal response regions as a function of both the location of the burst and the vowel context is given in Fig. 1.2 on the next page. As illustrated, the most important finding was that, not only does perception of place of articulation depend on the center

[5] because the upper frequency limit of Potter *et al.*'s spectrograph was 3500 Hz, it was impossible for them to observe the spectral peaks associated with alveolar releases as mentioned in Sec. 1.2.1

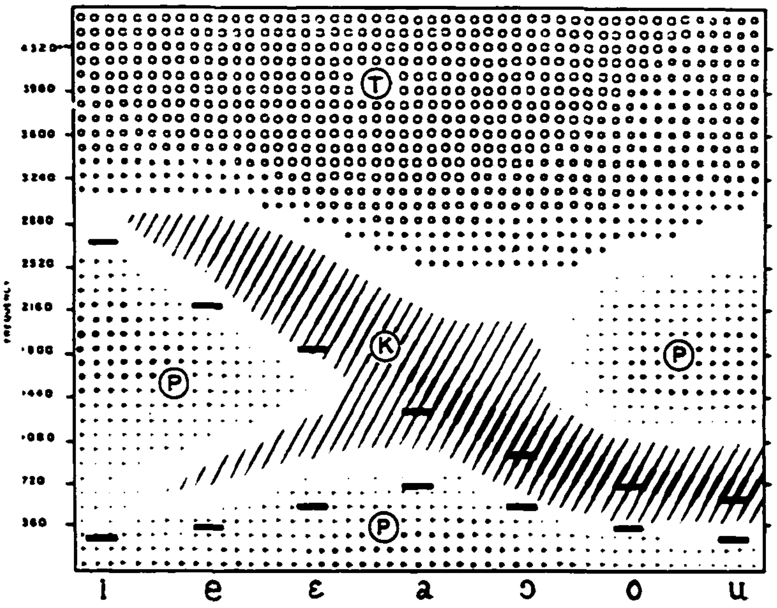


Figure 1.2: Modal decision regions based on responses to synthetic burst stimuli found in seven vowel contexts (Liberman *et al.*, 1952). From *American Journal of Psychology*. Copyright 1955 by the Board of Trustees of the University of Illinois. Used with permission of the University of Illinois Press.

frequency of the burst, but also on the context in which it occurs. When the burst was centered above 3 kHz, the stimuli were usually perceived as [t]. Below this value however, the tokens were heard as either [k] or [p] depending on the vowel that followed: [k] when the burst was just above the following F_2, otherwise [p]. In a similar experiment using a modern parallel formant synthesizer, Ainsworth (1968) found that the perception of the release depended on the resonance circuit used to produce the burst: [p] for F_1, [k] for F_2, and [t] for F_3 or F_4. These results are entirely compatible with the earlier findings. A consequence of the perceptual results obtained by Liberman *et al.* (1952) was that the same burst was heard as different consonants before different vowels—*e.g.*, a burst centered at 1440 Hz was perceived as [k] before [ɑ], but as [p] before [i] and [u].

This latter observation was verified by Schatz (1954) who cross-spliced bursts and vowels from different contexts using naturally produced speech. She found that the burst excised from [k^{h}ɑ] was heard as [p] when prepended to the vowels [i] and [u], but as [k] before the original vowel [ɑ]. While these results could not not be reproduced in a similar experiment by Cole and Scott (1974a), the latter authors included the aspiration portion with the burst thereby providing listeners with additional information from the voiceless formant transitions (see Section 1.3.2.1). Schatz, on the other hand, prepended the bursts to the syllables [^{h}i] and [^{h}u] which were produced by deleting the first few milliseconds of voiceless resonance from the syllables [hi] and [hu] respectively, in order to approximate the acoustic effects of aspiration. This would then be expected to have flat formant transitions similar to those synthesized by Liberman *et al.* (1952).

Based on what is now known regarding the spectral properties of stop bursts, the synthetic stimuli generated by Liberman *et al.* (1952) and Ainsworth (1968) can only be considered very crude approximations to the spectral patterns observed from detailed burst analyses. For example, Fischer-Jørgensen (1954) noted that release bursts contained multiple spectral peaks. More recently, several other researchers have given descriptions of burst spectral properties, a summary of which is given in Table 1.1 on the following page. Halle, Hughes, and Radley (1957) analyzed the first 20 ms of voiced and voiceless release bursts and found that the dominant concentrations of energy for the three places of articulation occupied complimentary spectral regions. Winitz, Scheib, and Reeds (1972) measured a single peak frequency in the spectra of pre- and postvocalic voiceless stops and provided absolute ranges for these values. Although the measurements are similar to those reported in other studies, it is most likely that they included the aspiration phase in their analysis as they report burst durations of up to 136 ms for [k] (see Section 1.2.1). Zue (1976) also measured the burst peak frequency from the first 10–15 ms of voiced and voiceless prevocalic stop bursts. In addition to finding a contextual effect for velar bursts, Zue noted that the distributions of [t] and [d] bursts were also bimodal: before the rhotic vowel [ɚ] and the rounded vowels [ɔ], [u], and [ʊ], mean peak frequencies were consistently lower than before unrounded vowels. It was also shown that, although velar bursts have a narrow spectral peak in the low-to-mid frequency region, a secondary peak was observed in the high frequency region before back vowels. In slight contrast with Winitz *et al.* and Zue, who measured only one spectral peak in their analyses, Jongman and Miller (1991) explicitly measured the two most prominent peaks for each burst spectrum and found that this resulted in better category separability in an ASR task (not shown in Table 1.1). Dorman *et al.* (1977) described the bursts from prevocalic voiced stops in terms of broad regions of relatively high spectral energy. Along with Halle *et al.* and Zue, they found that velar bursts were relatively compact with most of their energy occupying a narrow frequency band. All of these authors, with the exception of Winitz *et al.* and Jongman and Miller, found that velar bursts were highly variable and largely depended on the identity of the following vowel—*i.e.*, spectral energy was much lower in frequency before back vowels than before front vowels.

All of the above mentioned authors treated the oral release and the frication phases as a single acoustic unit—it is even suspected that Winitz *et al.* (1972) included the aspiration phase of voiceless stops as part of the burst. Only one experiment has attempted to analyze the spectral properties of the release independently of frication by recording very quietly whispered prevocalic stops (Repp and Lin, 1989). Repp and Lin found that, although oral releases showed very clear and distinctive formant-like structure, they were not found to be distinguishable in naturally produced bursts whose spectra were almost completely dominated by the acoustic energy of the frication phase. It was concluded therefore, that release transients were unlikely to have any perceptual significance.

1.3.2 Perception of the formant transitions

The observations made by Potter *et al.* (1947) regarding place information in the formant transitions were soon followed with perceptual experiments involving synthetic stimuli as well as more detailed analysis of spectrographic information. Similar to the

Author	Spectral energy		
	Bilabial (Hz)	Alveolar (Hz)	Velar (Hz)
Halle, Hughes, and Radley (1957)	500–1500	>4000	1500–4000[a]
Winitz, Scheib, and Reeds (1972)	1000–3350	3250–5750	1100–4200
Zue (1976)	...[b]	>3000	<3000[c]
Dorman, Studdert-Kennedy, and Raphael (1977)	<2000	>2000	...[d]

Table 1.1: Estimated distributions for spectral concentrations of release bursts.

[a] Halle *et al.* found that velar bursts showed higher frequency energy before front vowels [i] and [ɪ] (2–4 kHz) than before back vowels [ʌ], [ɑ], and [u]

[b] Zue found that bilabial bursts were too weak to be measured or otherwise showed no spectral prominences

[c] velar burst peaks were found to cluster according to three different vowel contexts: values were found at 2720 Hz, 1770 Hz, and 1250 Hz for front ([i], [ɪ], [e], [ɛ], and [æ]), back unrounded ([ɑ] and [ʌ]), and back rounded vowels ([ɔ], [o], [ʊ], and [u], and also including [ɚ]) respectively.

[d] close to the F_3 of a following front vowel or close to the F_2 of a following back vowel

work of Potter *et al.*, much of this research was focused on the second formant transition (Sec. 1.3.2.1). The first formant was generally thought to cue the distinction between voiced and voiceless stops (Sec. 1.3.2.2). However, it was later discovered that F_3 may also be used to discriminate place of articulation as well (Sec. 1.3.2.3). Section 1.3.2.4 describes some of the practical difficulties in measuring formant transitions from naturally produced speech which may have implications for their usefulness in modeling perception.

1.3.2.1 "Locus theory" and the perception of second formant transitions

The Pattern Playback was also used to investigate the perception of second formant transitions in *burstless* stimuli (Cooper *et al.*, 1952; Liberman *et al.*, 1954). The direction and extent of F_2 was varied before each of the seven two-formant vowels used previously in the perception of release bursts (Liberman *et al.*, 1952, Sec 1.3.1). Although it was assumed that the presence of the release burst was a major perceptual cue to voicing in prevocalic stops, it was found that the F_1 transition could also be used to indicate this distinction (Sec. 1.3.2.2) and subjects were therefore asked to identify the initial consonant as either [p], [t], or [k] when F_1 transitions were flat or as [b], [d], or [g] when F_1 was rising. Similar to findings regarding the perception of release bursts (Sec. 1.3.1), Cooper *et al.* and Liberman *et al.* found that the identification of place of articulation from second formant transitions also depended on the vowel context: while a rising F_2 was generally heard as bilabial, a falling F_2 was identified as either alveolar or velar depending on the following vowel. Another important finding was that the pattern of responses was very similar for both voiced and voiceless stops. In addition, it was also found that these same formant transitions similarly cued the place of articulation of the syllable-final nasals [m], [n], and [ŋ] indicating that place perception was independent of both manner and voicing distinctions. Similar results for the perception of second formant transitions were obtained by Ainsworth (1968).

A question remained regarding how the perceptual ambiguity of falling F_2 transitions could be resolved by listeners in naturally produced speech. It was suggested by

Cooper *et al.* (1952) that the spectral properties of the initial burst might be used in conjunction with formant transitions as *distinctive features* in a binary decision process similar to that proposed by Jakobson, Fant, and Halle (1965)—*e.g.*, high frequency burst indicated alveolar place, rising second formant transitions indicated a bilabial consonant, and a falling F_2 combined with a low frequency burst would otherwise be velar. However, Cooper *et al.* were not confident that this was the case.

Instead, it was noted that *flat* second formant transitions cued the perception of alveolar place only before [e] and [ɛ]—otherwise, only falling transitions cued this place of articulation. On the basis of this observation, Cooper *et al.* (1952) suggested that F_2 transitions in prevocalic alveolar stops might point to a characteristic "locus". Although the locus concept was not directly borrowed from Potter *et al.* (1947), it was in principle functionally equivalent to a hub: a point of convergence for F_2 in different vowel contexts for a single place of articulation. Although they should have expected that a rising F_2 transition before [i] would also have been heard primarily as [d] or [t], Cooper *et al.* and Liberman *et al.* were unable to elicit a majority of alveolar responses to *any* formant transition before this vowel, presumably because a rising F_2 in this context was normally heard as either [b] or [p] in the absence of a high frequency release burst (Section 1.3.1)[6].

The following line of reasoning was then put forth by Delattre *et al.* (1955). It was known that formant transitions reflected changes in the size and shape of the vocal tract and that these changes resulted primarily from the movement of the oral articulators from the place of articulation of the consonant to that of the vowel. It was then suggested that the onset of the formant transitions gave the listener insight into the approximate configuration of the speaker's vocal tract at the exact moment immediately following release. Because it was assumed that *articulation* did not in general vary with vocalic context, it was concluded that the spectral properties at the onset of formant transitions—in particular that of F_2—should reflect this articulatory invariance and therefore provide the primary acoustic source of place of articulation information. This was labelled as the "simple" locus hypothesis: that each place of articulation had a perceptually relevant, distinctive locus. They found that the locus for alveolars was most evident from perceptual experiments (Cooper *et al.*, 1952; Liberman *et al.*, 1954) in which flat formant stimuli before both [e] and [ɛ] generated the most [t] and [d] responses (*cf.*, Potter *et al.*, 1947; Fischer-Jørgensen, 1954). By similarly studying the perception of *flat formant* stimuli in different vowel contexts, Delattre *et al.* concluded that the loci for alveolars and bilabials were 1800 Hz and 720 Hz respectively. Contrary to the observations of Potter *et al.* (1947), who found F_2 *always* falling after [g] and [k], Delattre *et al.* concluded that velars also possessed a fixed locus at 3000 Hz. This estimate appeared to be unsatisfactory before the back vowels, which already had very low F_2's, and therefore no single unique locus was found for [k] and [g]. Nevertheless, Delattre *et al.* were undeterred from their assumptions regarding the relationship between formant transitions and articulation and the contextual invariance of articulation itself.

However, when second formant transitions were synthesized to actually originate at the proposed alveolar locus, the stimuli were instead heard either as [b], [d], or [g]—once again depending on the F_2 of the following vowel. Alveolars were not consistently heard until the first half of the formant transition was deleted, indicating that

[6] it is also interesting to note that Liberman *et al.* (1952) were unable to elicit a majority of [ki] responses in the burst experiment described in Sec. 1.3.1, presumable because a high frequency burst was heard as [t] in the absence of a falling second formant transition

the second formant must point to, but not actually reach its characteristic locus. This was presumably because, in natural speech, the frication phase obscures much of the formant transition that would otherwise be present. Similar results were obtained for bilabial transitions (Delattre *et al.*, 1955).

Based on an acoustic theory of speech production, we should expect a relatively low-frequency bilabial locus. If the human vocal tract is approximated by a uniform tube, open at one end (at the lips) and closed at the other (the glottis), we expect to see resonances at odd multiples of the first formant. For example, a 17.5 cm uniform tube resonates at approximately 500 Hz, 1500 Hz, 2500 Hz, *etc.* In this example, when the acoustic resonator is closed at the opening, the second formant frequency is lowered to approximately 1000 Hz.

The relationship between articulation and formant transitions was tested empirically by Stevens and House (1956) using an electrical analogue of the human vocal tract to synthesize speech. It was found that the locus of velar consonants—and to some extent bilabial consonants as well—moved in response to the identity of the following vowel and did not necessarily reflect the assumed articulatory invariance in these cases. However, it was found that bilabial loci were always lower than the F_2 of the following vowel and that velar loci were generally higher, resulting in consistently rising and falling transitions respectively. In an examination of naturally produced prevocalic formant transitions, Kewley-Port (1982) found that bilabial loci, which were extrapolated backwards in time from the F_2 transition to the onset of the release, formed two clusters: one for the front vowels [i], [ɪ], [e], [ɛ], and [æ], and another for the back vowels [ɑ], [o], and [u]. Distinct bilabial loci for these two contexts were estimated at 1645 Hz and 1090 Hz respectively. No clustering was found for velar loci at all, while a single alveolar locus was estimated at 1797 Hz.

The notion of a characteristic locus for each place of articulation has had a profound impact on theories of speech perception in general. It ultimately provided the basis for the *Motor Theory* of speech perception (Sec. 1.4.1) and also led to the formulation of *locus equations* as semi-invariant place-of-articulation cues as presented by Sussman (1991) and Sussman, McCaffrey, and Matthews (1991).

To complicate matters however, Öhman (1966) found that formant transitions following *intervocalic* VCV Swedish stops depended not only on the following vowel, but also on the preceding one. For example, F_2 falls by 205 Hz following the burst in [ybɑ], but rises by 100 Hz in [obɑ] even though the *following* vowel context [ɑ] is identical. It was also found that there was a considerable amount of overlap in measured loci for [b] and [d]. On the basis of x ray measurements, it was concluded that much of this contextual variation was due to coarticulation between the medial consonant and both the preceding and following vowel. Similar observations were made by Fant (1973).

However, in a followup study using similar VCV syllables in French, Delattre (1969) concluded that the only stimuli in which F_2 failed to point to a characteristic locus were those containing [g] adjacent to a rounded vowel and that these stimuli also showed a strong characteristic burst following the silent phase. He suggested that the variability of velar formant transitions was compensated by the strength of the release cue instead and that the perception of *burstless* stimuli was best understood with reference to the locus theory of speech perception (Sec. 1.3.4.3). This complimentary relationship between burst and formant transitions was also proposed by Dorman *et al.* (1977) who suggested that it was generally the case that when one cue was perceptually weak, the other was relatively strong (Sec. 1.3.4).

1.3.2.2 First formant transitions

In contrast with the results regarding the perception of F_2 transitions, first formant transitions were not generally found to affect the perception of place of articulation. In early experiments that studied these cues, Cooper *et al.* (1952) and Liberman *et al.* (1954) used F_1 to cue the perception of voicing contrasts: a rising F_1 was generally perceived as one of the voiced stops [b], [d], or [g], while a flat F_1 was heard as one of the voiceless stops [p], [t], or [k].

Ultimately, Delattre *et al.* (1955) declared that F_1 had no influence on place perception, and it was further proposed that all voiced stops had a single F_1 locus somewhere below 240 Hz. This was supported by observations made by Fischer-Jørgensen (1954) who found that voiceless stops showed a straight F_1 while voiced stops clearly showed a rising first formant, especially before [a], which already had a relatively high-frequency F_1. Based on an acoustic theory of speech production, we expect the F_1 locus for all stops to be quite low, since closure anywhere within the oral cavity will result in infinite acoustic impedance of the vocal tract behind the constriction, or $A/l = 0$, where A is the area of the opening at the point of constriction, l is the distance from the point of constriction to the mouth opening, and A/l is a measure of acoustic conductivity. The limit of F_1 frequency as $A \rightarrow 0$ is 0 Hz and is also therefore independent of the constriction location l (Stevens and House, 1956; Fant, 1970, p. 63). Nevertheless, we do expect differences in the *rate of change* in F_1 as a function of constriction location. In fact, Stevens, Manuel, and Matthies (1999) have suggested that since F_1 should have a more rapid transition for bilabials, this property may additionally serve to discriminate place of articulation.

The absence of a rising F_1 transition in voiceless stops is likely due to the presence of aspiration which obscures its onset. It is believed that the open glottis during aspiration introduces antiresonances that largely cancel out this formant. Liberman, Delattre, and Cooper (1958) found that perception of voicing depended not only on the onset frequency of F_1, but also on its latency relative to the higher formants, referred to as F_1-cutback.

1.3.2.3 Third formant transitions

Because most of the Pattern Playback synthesis experiments focused on two-formant stimuli, F_3 was largely ignored as a possible perceptual correlate to place of articulation. Prior to these experiments, it had already been observed that the second and third formants formed an "egg shape" in velar transitions, in which F_2 and F_3 seemed to originate from a single hub or locus (Potter *et al.*, 1947). This was particularly noticeable before the vowel [æ]. Nothing was however mentioned regarding the pattern of F_3 transitions in other stop consonants.

In informal listening tests however, Liberman *et al.* (1954) found that a falling F_3 enhanced the perception of alveolar place of articulation, while a rising F_3 improved the perceptual quality of both bilabials and velars depending on whether the second formant was rising or falling respectively. It was thought that this additional acoustic cue might further disambiguate the perception of falling F_2's in burstless, synthetic stimuli.

Fischer-Jørgensen (1954) estimated F_3 loci for the three places of articulation in a similar manner as had been done for second formant transitions (Section 1.3.2.1). For bilabials and alveolars, she found F_3 loci at 2300 Hz and 2700–2800 Hz respectively.

However, no F_3 locus could be estimated for velars; she found that F_3 was level before front vowels, and rising before back vowels. She also noted the same characteristic convergence of the second and third formants following velar bursts, but only before the Danish vowels [a], [y], [ø], and [œ]. Similar results were obtained by Delattre (1969).

An empirical perceptual experiment using synthetic stimuli was eventually designed by Harris *et al.* (1958) who used the Pattern Playback to manipulate second and third formant transitions in three-formant stimuli in only two vowel contexts: [i] and [æ] where F_3 was synthesized at 3000 Hz and 2400 Hz respectively. Liberman *et al.*'s (1954) suspicions regarding the potential role of third formant transitions were confirmed: falling (and to a lesser extent, straight) F_3's increased the number of [d] responses, while rising F_3's before [æ] enhanced the perception of [b] and [g] when found in conjunction with the appropriate second formant transition. No improvement in [g] responses was ever found before [i] however.

Based on Fischer-Jørgensen's (1954) measurements, these results were expected. Measurements made by Öhman (1966) on Swedish F_3 transitions also support these results. For example, F_3 was found to be always rising following [g] and [b] and always falling after [d] when the initial vowel in the VCV sequence was [o]. Lindblom (1990) showed that the addition of onset F_3 frequency to the frequencies of F_2 at the onset and vowel steady-state portions increased the separability between the observed distributions of the three places of articulation such that they were no longer overlapping. Therefore, he claimed that the observed variance *in F_2 loci found in Öhman's study* was actually the result of natural covariation which could be factored out by taking into account additional acoustic parameters such as F_3.

1.3.2.4 Problems with measurement of formant transitions

Although evidence of the perceptual importance of formant transitions on the basis of synthetic speech experiments seems compelling, the measurement of actual formant onsets and steady states from spectrograms of real speech presents significant difficulties. This creates a particular problem for researchers who wish to assess the perceptual importance of these features from naturally produced stimuli. For example, much of the locus theory presupposes that the listener is able to determine the steady state formant frequencies of the vowel following the syllable-initial consonant. However, very few vowels in English exhibit any type of steady state behavior. While this is particularly true for very short vowels, such as [ə] and [ʌ] or diphthongs such as [aɪ] and [aʊ], it is also true for many other vowels that have traditionally been labelled as monophthongs such as [ɛ] and [ɪ]. Nevertheless, many researchers have measured the extent and duration of consonant formant transitions under the false assumption that many English vowel formants are static over some period of time. For example, Lehiste and Peterson (1961) assumed that lax vowel formants, such as those found for [ɛ] and [ɪ], possess some kind of static formant pattern and that good estimates of formant transition duration can be obtained in these vowel contexts. However, it is now known that these vowels show just as much dynamic formant frequency movement as do vowels such as [e], which are traditionally labelled as diphthongs in English (Nearey and Assmann, 1986; Hillenbrand *et al.*, 1995).

This problem has occasionally been solved by avoiding reference to the "steady state" portion of the vowel. For example, LaRiviere, Winitz, and Herriman (1975) determined vowel formant information from the moment of maximum amplitude of the

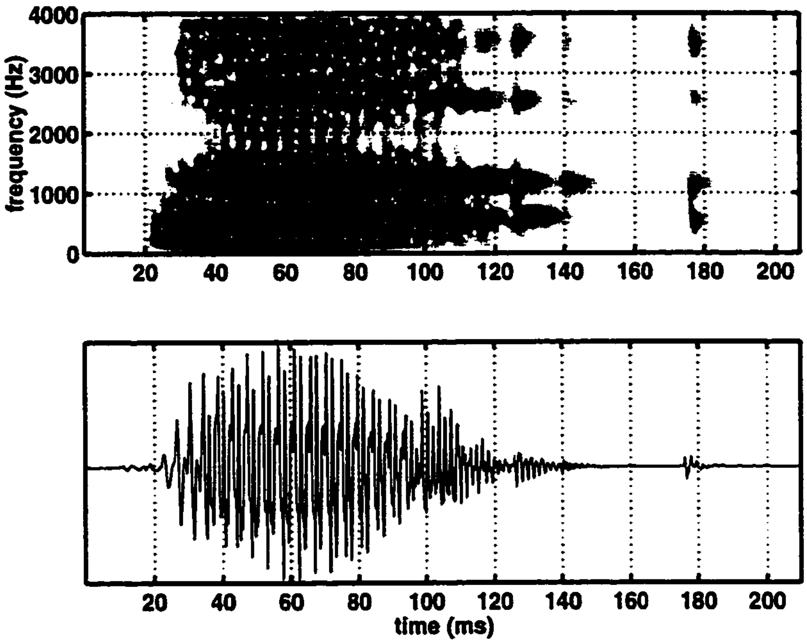


Figure 1.3: Example spectrogram showing differing onset times for F_1 and F_2. The wideband spectrogram is of the word [buk] ("book") as produced by a female speaker. Analysis parameters are the same as those for Fig. 1.1 on p. 6.

waveform. Ohde and Sharf (1977) analyzed the 40 ms segment of the vowel which showed the *least* amount of formant frequency change. Shammass (1985) defined the vowel steady state as the point 60 ms following the onset of vocalic formants provided that these frequency values were within 50 Hz of measurements taken from adjacent analysis windows. Schouten and Pols (1983) suggested that the most important transitional information relating to the identity of the initial stop is contained within the first three pitch periods following the onset of voicing. However there are additional problems in measuring the *onsets* of the formants themselves.

It is often very difficult to detect the presence of formant resonance in the aspiration phase of voiceless stops and many authors have even associated this portion with the preceding burst, opting instead to measure formant onset at the beginning of the voiced segment (Sec. 1.2.1). It is also known that F_1 does not begin simultaneously with the higher formants in voiceless stops due to F_1-cutback (Sec. 1.3.2.2). However, it is also possible for the onset of F_1 to begin before the remaining formants, especially in voiced stops were the first few glottal pulses or "edge vibrations" (Lisker and Abramson, 1964) are near sinusoidal and excite only the lowest frequency regions. Therefore, decisions regarding when formant onsets actually occur is not always so clear, especially when they do not all begin simultaneously. Figure 1.3 illustrates a case in which F_1 begins at least one full glottal period before F_2 or F_3. In addition, F_2 appears to begin before F_3. Note that the first glottal pulse (immediately following the 20 ms grid line) is very nearly sinusoidal.

It should be noted that all of these concerns apply equally well to *locus equa-*

tions (Sussman *et al.*, 1998). Although Sussman (1991) provides complicated criteria for locating the vowel nucleus, much of these measurements rely on the visual estimation of the vocalic midpoint from spectrograms.

1.3.3 Other acoustic properties

Although the spectral properties of the burst and formant transitions have formed the core set of proposed detailed cues, purely *temporal* features are potentially important as well. Ohde and Stevens (1983) found that the amplitude of the burst relative to that of the following vowel has an effect on the discrimination between [p] and [t] with louder bursts favoring [t]. Similarly, it was suggested by Cole and Scott (1974b) that discrimination between [p] and [k] could be done on the basis of the waveform envelope alone.

Although Cole and Scott did not explain how this might be done, this distinction may relate to differing VOTs between places of articulation as well as the presence of multiple release bursts for velars: the length of the release burst, as well as VOT, have been shown to be somewhat distinctive between different places of articulation. For example, Fischer-Jørgensen (1954) found that [g] bursts were much longer than those of alveolar and bilabial stops and that they often showed double clicks or multiple releases with intervening 5–10 ms silent gaps. This was also illustrated by Dorman *et al.* (1977). That velars have longer VOTs than other places of articulation was further supported in an extensive cross-language study of voice onset time (Lisker and Abramson, 1964). Kewley-Port (1982) claimed that she was able to determine the place of articulation of prevocalic stops, based solely on VOT information, with 88% accuracy. Although this result has never been replicated, it has nevertheless been shown in perceptual experiments involving synthetic continua, that VOT is indeed a secondary cue to place of articulation. For example, Oden and Massaro (1978) have shown that there is a change in the response boundary between alveolars and bilabials along an F_2–F_3 continuum as a function of VOT. Similarly, Benkí (1998) has shown that VOT has a significant effect on place judgments.

1.3.4 Relative importance of acoustic cues

Potter *et al.* (1947) relied on the spectral properties of the formant transitions in discriminating place of articulation from speech spectrograms. This was done mostly out of necessity; spectrograms are best suited for observing dynamic spectral patterns over periods of time much longer than the typical duration of bursts. Nevertheless, Joos (1948) claimed that the perception of stop place of articulation *must* be made primarily on the basis of formant transition information, since isolated bursts cut from tape recordings of natural speech revealed little about their identity, other than their manner or voicing. Similar to Potter *et al.*, it has been believed by many that the second formant alone contains most of the information relevant to speech perception. Liberman *et al.* (1967, p. 434) state:

> The second formant transition is a major cue for all the consonants except, perhaps, the fricatives /s/ and /sh/ and probably the single most important carrier of linguistic information in the speech signal.

However, it is clear that the release burst contains much relevant place information as well (Sec. 1.3.1) and it is even believed that the relative perceptual weights of the burst and the formant transitions are in some way complimentary (Delattre, 1969; Dorman *et al.*, 1977, see also Sec. 1.3.2.1). For detailed cue theories of speech perception where a distinction is made between the perceptual properties of the burst and formant transitions, it is important to establish how such disparate cues are integrated to form a single percept. Several researchers have attempted to answer this question by evaluating the relative perceptual importance of burst and transition cues by deleting either one or the other from naturally produced stimuli which are then presenting to listeners for categorization. Another approach has been to splice together bursts and formant transitions from conflicting places of articulation to evaluate which set of cues dominates perception. These experiments are described in detail in the following sections.

The most straightforward approach in evaluating the perceptual integration of multiple cues has been to parametrically manipulate the spectral properties of several such cues simultaneously in synthetically produced stimuli. Such experiments are described in Sec. 1.3.4.1. Sections 1.3.4.2 and 1.3.4.3 present data from experiments in which burst-only and burstless stimuli are presented to listeners for identification. Section 1.3.4.4 describes experiments using mixed cue or conflicting cue stimuli.

1.3.4.1 Parametric manipulation of multiple cues

Only two experiments involving the use of synthetic stimuli have attempted to manipulate the acoustic properties of both the release burst and formant transitions. However, one of them (Ainsworth, 1968) provided only two levels of formant transitions for each of F_2 and F_3. Ainsworth synthesized the burst using one of four available resonators thereby extending one of these formants backwards into the initial, noise excited portion. Despite the fact that the release burst and formant transitions were spectrally correlated, he assumed from the outset that the parameters were *statistically independent* and pooled results within conditions. Based on such data, he made the claim that the release burst was more important for [g]—*i.e.*, the direction of either F_2 or F_3 formant transition had relatively little effect on its perception. In addition, he noted that [d] bursts were more important before the back vowels [ɔ], [ɑ], [o], [u], and [ʊ] than before the front vowels [æ], [ɛ], [ɪ], and [i]. This relationship appeared to be true for the voiceless stops as well. Otherwise, the results were very similar to those found elsewhere (*e.g.*, Liberman *et al.*, 1952, 1954; Cooper *et al.*, 1952).

In a more detailed synthesis experiment, Hoffman (1958) manipulated the second and third formant transitions independently of the release burst, which was otherwise similar to the experiment performed by Liberman *et al.* (1952) (Sec. 1.3.1). However, in the interest of keeping the total stimulus set small, Hoffman used only one vowel context [æ]. He also included burstless and two-formant stimuli to test the relative importance of these cues. The results did not in general conflict with other findings: rising F_3's increased [b] and [g] responses while a falling third formant increased [d] responses (*cf.*, Harris *et al.*, 1958, Sec. 1.3.2.3). In addition, he found that high-frequency bursts increased the number of [d] responses while low-frequency bursts increased [g] responses. Unlike the results obtained by Liberman *et al.* (1952) however, Hoffman could find no burst that improved the perception of [b] above and beyond what was found with the burstless stimuli. This indicated that the release burst was perhaps a relatively unimportant acoustic feature for this consonant. Another important conclusion drawn from this experiment was that the addition of either a release burst or third formant affected the pattern of responses in roughly the same way whether the third formant or the burst was already present respectively. This indicated that their perceptual effects were relatively independent.

1.3.4.2 Perception of isolated release bursts

Several experiments have assessed the sufficiency of the burst cue from naturally produced speech when presented in isolation to listeners. Halle *et al.* (1957) segmented the first 20 ms of the *syllable final* release of the voiceless stops [p], [t], and [k] to listeners for identification and found that subjects who had had minimal experience with such stimuli were able to correctly identify the place of articulation with 65–70% accuracy. Listeners with considerably more experience in listening to such stimuli performed much better (75–96% correct identification). A similar experiment was performed by Malécot (1958) using the final releases of French voiced stops taken from two vowel contexts, [ɔ], and [ɛ]. Listeners in this experiment were able to identify place information from these stimuli with an average of 85% accuracy. In both cases, chance performance was 33%. Syllable final releases were particularly appealing in isolated burst experiments simply because it was much easier to segment them from natural speech; their onset is clearly bounded by the preceding silence of the closure phase. While the onset of release bursts from prevocalic stops is easily located, it is often difficult to decide where the frication energy ends and the aspiration portion begins—especially for voiceless stops and velars. Since the latter often show a compact burst peak contiguous with the following F_2, it is easily confused with the following formants.

Despite these difficulties, several other experiments have addressed the perception of isolated, syllable-initial release bursts. For example, Schouten and Pols (1983) found 65.7% correct identification for voiceless stops taken from the onsets of Dutch syllables (short-lag VOT) as compared with 33.3% for chance levels of performance. Although they determine identification rates for voiced stops as well, Dutch does not have the consonant [g], which makes comparison between the two scores rather difficult.

In a similar isolated burst perception experiment performed by Kewley-Port, Pisoni, and Studdert-Kennedy (1983), it was found that [b] bursts were correctly identified 90% of the time. Voiced alveolar bursts were identified at similar levels of accuracy. The burst excised from [g] however, was identified quite poorly with only 70% correct responses. This is quite contrary to the results obtained by Smits *et al.* (1996a) who obtained correct identification rates of 80%, 49.6%, and 91.1% for isolated Dutch [p], [t], and [k] bursts respectively (73.6% overall). Notably, correct recognition from bursts excised from [ka], [ky], and [ku] were above 95%. Similarly, Bonneau, Djezzar, and Laprie (1996) obtained 76%, 91%, and 94% for [p], [t], and [k] respectively, in a task involving segmented French bursts. However, these rates were reduced to 89%, 86%, and 87% when burst durations for the isolated segments were normalized, thereby eliminating duration information. In all of these cases, chance performance would be indicated by 33% correct recognition. The discrepancy between these latter results and those of Kewley-Port *et al.* may be due to the difference in segmentation procedure for the bilabial stop: Kewley-Port *et al.* included the first glottal pulse following the burst for [b]. This was justified on the basis that voicing was often continuous throughout the duration of the bilabial release, which was already weak in intensity, and therefore difficult to isolate. The results may also conflict because of additional language specific differences, even though VOTs for the three sets of stimuli are very similar.

It is typically held that neither French nor Dutch voiceless stops contain an aspiration segment. Therefore, the problem of whether to include such a segment did not present itself—likewise for voiced stops in English (*e.g.*, Kewley-Port *et al.*, 1983).

Although Winitz *et al.* (1972), LaRiviere *et al.* (1975) and Ohde and Sharf (1977) give results on the perception of isolated *voiceless* bursts in English, their stimuli included the aspiration segment and therefore contained voiceless formant transitions. It was later shown (Just *et al.*, 1978; Pols and Schouten, 1978) that the presence of aspiration in voiceless bursts substantially improves the perception of place of articulation and so these results cannot be interpreted on the basis of the burst properties alone.

Tekieli and Cullinan (1979) presented listeners with the first 10–150 ms of CV syllables containing the voiced and voiceless stops as well as the affricates [ʤ] and [ʧ] in eight vowel contexts: [i], [ɪ], [u], [ʊ], [æ], [ɛ], [ɑ], and [ə]. Although they did not isolate the burst from these consonants explicitly, they did find that place of articulation was correctly identified 60% of the time when only the first 10 ms was presented and 90% of the time for the first 20 ms (VOTs for [b], [d], and [g] were 26 ms, 28 ms, and 34 ms respectively). It would appear then, that very good place identification can be obtained from very short segments taken from the onsets of prevocalic stops. Nearey (1992c) found similar rates of correct identification with up to 80% for 13 ms stimuli and 70–90% for 18 ms stimuli. Krull (1990) found that 70.6% of stimuli segmented from the first 24–27 ms following the oral release of Swedish voiced stops were identified correctly for place of articulation. Although Swedish possesses four possible places of articulation for stops including retroflex [ɖ], pooling dental [d̪] and retroflex responses resulted in a correct identification rate of 79.4%.

In the above experiments, the vocalic context was not included in the presentation stimuli despite the fact that this has been shown to be a very important factor in the perception of the release burst (Liberman *et al.*, 1952). Fischer-Jørgensen (1972), Cole and Scott (1974a), Dorman *et al.* (1977), and Just *et al.* (1978) presented listeners with stimuli in which only the *formant transitions* had been deleted with the burst and the steady state vowel context remaining. Dorman *et al.* substituted a silent portion for the original formant transitions, while the bursts in the remaining experiments were joined with the steady state portion of the vowel. Fischer-Jørgensen found that correct identification of voiced place of articulation was poor before [a], but otherwise relatively good before [i] and [u] (24%, 77%, and 69% respectively). An examination of the confusion matrices given in the paper reveals, however, that most of the errors found in the [a] context were due to listeners' inability to even detect the presence of a stop: *e.g.*, 82% of transitionless [ba] stimuli were heard as having no stop at all. This is not surprising given that voiced bilabial bursts have very little energy. In contrast Cole and Scott obtained almost perfect identification performance on similar stimuli (99.5% correct). However, they used fixed durations when determining the boundary between burst and formant transitions: *i.e.*, 20 ms, 30 ms, and 40 ms for [b], [d], and [g] respectively—these appear to be about 10 ms longer than typical burst length measurements (*cf.*, Tekieli and Cullinan, 1979). Dorman *et al.* noted that bilabial bursts were relatively weak perceptual cues. Alveolar bursts were found to be most effective before the front vowels [i], [ɪ], and [ɛ] but weak when found in the context of vowel rounding. Velar bursts were either found to be very weak cues or were at best, most effective when found before back, rounded vowels. However, it was noted that there were substantial speaker differences in this study and the fact that there were only two speakers makes these differences difficult to evaluate. Just *et al.* found results similar to those of Fischer-Jørgensen: namely stop consonant identification from transitionless stimuli was better in the [i] and [u] contexts than before [a] (93%, 93%, and 66% respectively). Just *et al.* also found that there was a significant inverse correlation between percent correct identification and the duration of the formant transitions which were found to

Author	Rate (%)	Stimuli	Context
Sharf and Hemeyer (1972)	67.7[a]	English vcd.	[ə]
Fischer-Jørgensen (1972)	46.4[a]	Danish vcd.	[i,a,u]
Ostreicher and Sharf (1976)	49.3	English vcd.	[i,ɚ,ɔ,o,u]
Ohde and Sharf (1977)	44.0	English vcd.	[i,ɚ,u]
Pols and Schouten (1978)	55.2	Dutch vcls.	[i,a,u]
(masking noise)	73.3		
Pols (1979)	60.0	English vcd.	[i,ɑ,u]
(masking noise)	83.0		
Ohde and Sharf (1981)	72.0	English vcd.	[i,e,ɑ,o,u]
Schouten and Pols (1983)	46.8	Dutch vcls.	[i,ɪ,y,ʏ,e,ɛ,ø,a,ɑ, o,ɔ,u,ɛi,œy,əu]
Smits, ten Bosch, and Collier (1996a)	68.8	Dutch vcls.	[i,y,a,u]

Table 1.2: Correct identification rates for burstless stimuli.

[a] Subjects were also permitted to respond with "no consonant"

be longest before [ɑ]. Although most of the errors were in place identification, these results were in fact pooled over place and voicing judgments. In a later study, Schouten and Pols (1983) found that 68.5% of transitionless syllable-initial Dutch voiceless stops were identified correctly for place of articulation.

The only general statement that can be made regarding this data is that subjects were able to identify isolated bursts with a reasonable amount of accuracy. Aside from this, there appears to be very little consistency between the various studies. This may be due to a number of factors: language-specific differences—*i.e.*, English (*e.g.*, Kewley-Port *et al.*, 1983) *vs.* Dutch (*e.g.*, Smits *et al.*, 1996a) *vs.* French (Bonneau *et al.*, 1996, *e.g.*,), differences in segment cutting procedure [Kewley-Port *et al.* segmented bursts at zero-crossings, but did not further taper the signal with a windowing function as did Smits *et al.*], or differences in response options [*e.g.*, Fischer-Jørgensen (1972) allowed subjects to respond with "no consonant"].

1.3.4.3 Burstless stimuli

Several researchers have measured listeners' ability to detect the correct place of articulation from *burstless* stimuli. These results are summarized in Table 1.2. As was the case for burst-only stimuli, the results show a wide range of performance: from 44% in Ohde and Sharf's (1977) study to 83% found by Pols (1979) using a masking noise to replace the deleted burst.

Only two of these studies attempted to determine listeners' ability to detect the presence *vs.* the absence of the consonant from burstless stimuli (Sharf and Hemeyer, 1972; Fischer-Jørgensen, 1972). Fischer-Jørgensen found that when bursts were removed before [i] and [u], there was a strong tendency for listeners to respond with "no consonant", except in the syllables [bi] and [du] which have very steep, rising and falling second formant transitions respectively (Sec. 1.3.2.1). Ohde and Sharf (1977) also found that [bi] and [du] are identified correctly at very high rates (79% for both). However, Fischer-Jørgensen also found an overall strong bias towards [b] (45% of all responses), a tendency that has also been found elsewhere (*e.g.*, Smits *et al.*, 1996a).

In an additional study not presented in Table 1.2, Dorman *et al.* (1977) found that bilabial formant transitions were almost as effective a cue to place of articulation as whole CV syllables. This could again simply be a case of bilabial bias as found elsewhere—*i.e.*, if all responses are [b], this will result in nearly 100% correct identification for bilabial place of articulation. However, they also note that alveolar formant transitions were strong cues before back vowels, but were otherwise weak. Velar transitions were always found to be weak cues. Similar to Just *et al.* (1978, see also Sec. 1.3.4.2), they also made the suggestion that the effectiveness of formant transitions might be related to their duration.

Some of the poorer identification rates found in Table 1.2 on the preceding page may be related to the technique used to remove the burst from recorded speech. In many of these studies (Sharf and Hemeyer, 1972; Fischer-Jørgensen, 1972; Ostreicher and Sharf, 1976; Ohde and Sharf, 1977), the signals were segmented by manually cutting the tape on which the stimuli were recorded. It was argued by Pols and Schouten (1978) that this often leaves a residual "click" sensation that may bias the perceptual results towards bilabial responses (*cf.*, Fischer-Jørgensen, 1972). This problem could largely be circumvented by digitally cutting the signal at zero crossings (*e.g.*, Pols and Schouten, 1978; Fujimura *et al.*, 1978) which greatly improves the number of correct identifications. Pols (1979) also showed a significant improvement in stop identification when the deleted burst was instead replaced by 300 ms of pink, speech-shaped noise. More recently, it has been found that tapering the onset of the signal with an appropriate windowing function is just as effective (Ohde and Sharf, 1981; Schouten and Pols, 1983; Smits *et al.*, 1996a).

1.3.4.4 Perception of mixed cue stimuli

In an attempt to find the minimal number of acoustic segments that would be needed to produce intelligible speech from a concatenation of isolated speech sounds, Harris (1953) found that when alveolar and bilabial stop bursts were spliced onto velar formant transitions, listeners never responded with velar place of articulation. Aside from this very limited experiment, only two papers have investigated the perception of *mixed-cue* or *conflicting-cue* stimuli. Fischer-Jørgensen (1972) found that listeners' identification of place of articulation was appropriate to the burst in voiced stops before [i], while transitions were more decisive before [a]. Before [u], results were mixed. In a very similar experiment using Dutch stops, Smits *et al.* (1996a) found that the [k] burst was a significantly stronger cue than the bursts of either [t] or [p]. However, they could find no significant differences in the effectiveness of [p], [t], or [k] transitions. In terms of vowel contexts, the burst was most effective before [y], then [i] and [u] and least so for [a].

1.3.4.5 Summary

Despite many conflicting results, some conclusions can be drawn regarding the relative importance of burst and transition cues in the perception of prevocalic, voiced (or short-lag) stops. It has generally been found that bilabial bursts are not important for the perception of this place of articulation. For example, Hoffman (1958) found that the presence of a burst did not increase the number of responses to [b] regardless of its spectral properties. Likewise, it has been found that *burstless* stimuli are more often confused with bilabials than with any other place of articulation suggesting that even the *absence* of a release cue biases responses to this consonant (*e.g.*, Fischer-Jørgensen, 1972; Smits *et al.*, 1996a).

Conversely, it has been shown that velar bursts are very strong perceptual cues in place-of-articulation identification (*e.g.*, Smits *et al.*, 1996a). However, the importance of formant transitions appears to be less well understood. Nevertheless, it has been suggested that there is a complimentary relationship between the perceptual weight of formant transitions suggesting that bilabial formant transitions are much more important than those of velar stops (*e.g.*, Dorman *et al.*, 1977), despite the fact that this difference has not been shown empirically (*e.g.*, Smits *et al.*, 1996a).

1.4 Theories of speech perception

1.4.1 The Motor Theory of speech perception

In the search for invariant correlates to place of articulation in stop consonants, the results obtained in early experiments using synthetic speech led many researchers to adopt the position that perception was mediated in some way by articulation. This was largely inspired by the complex relationships observed between articulation and acoustics which ultimately resulted in simple discrete percepts such as different places of articulation for stop consonants. It was expected then that the relationship between *articulation* and *perception* would be relatively simple. For example, the action of pressure release by the tongue apex at the alveolar ridge was generally heard as [d] or [t] regardless of the vowel that followed it. This simple relationship was clearly not evident between *acoustics* and perception. Experiments showed that there was often a many-to-one or a one-to-many mapping between a perceived consonant and the acoustic stimulus parameters that cue it. For example, it was observed that the same burst centered at 1400 Hz was heard as [p] before [i] and [u], but as [k] before [ɑ] (Liberman *et al.*, 1952; Schatz, 1954, see also Sec. 1.3.1). Conversely, different bursts for [k] were required before different vowels to generate a velar percept (Liberman *et al.*, 1952). The proposed simple relationship between perception and articulation was based on the context sensitivity of observed acoustic properties and the assumed context-invariance of articulation. It was therefore suggested that perceived similarities in the speech signal corresponded more closely to the articulatory domain.

Although it has been suggested that the movement of the oral articulators can be perceived directly [*e.g.*, Fowler's (1986) direct-realist approach to speech perception], Liberman *et al.* (1967) instead claimed that the invariant component of speech was somehow *encoded* into the acoustic signal in a relatively complex manner.

Several observations were used to illustrate the complexity of the acoustic signal relative to the articulations that produce it. Firstly, a stream of "building blocks", in which discrete acoustic segments represented unique units of speech, would be very difficult to understand at the rates needed for the perception of continuous speech (*cf.*, Harris, 1953); communication by such means would be no more efficient than Morse code. Subphonemic cues are in fact widely overlapping in natural speech. For example, we find acoustic properties relevant to the place of articulation of prevocalic stop consonants in the burst, formant transitions, and vowel formants, while vowel information can be found in the formant transitions and vowel formants. This lack of phoneme *segmentability* was seen as a consequence of the relative efficiency with which speech could be transmitted acoustically—*i.e.*, several speech units could be conveyed simultaneously at any given moment.

The second observation that illustrated the complexity of the acoustic signal relative to its articulatory source and its perceptual representation, was the nonlinear mapping between articulation and acoustics on one side, and the nonlinear mapping between acoustics and perception on the other. As has already been shown, very different stimuli are heard as the same speech unit: *e.g.*, although formant transitions for [di] and [du] are quite different (F_2 rising and falling respectively), they are nevertheless perceived as the same consonant.

Place of articulation has also been found to exhibit very nearly *categorical perception* in which listeners' ability to discriminate acoustically similar sounds depends critically on whether the two sounds are perceived as different phonemes. Liberman *et al.* (1957) and Eimas (1963) found that listeners' ability to discriminate stimulus continua was just slightly better than what could be predicted on the basis of place-of-articulation identification responses alone—*i.e.*, discrimination was much better *across* phoneme boundaries than within a single speech category. Eimas showed that this phenomenon was not found with nonspeech stimuli and was therefore peculiar to speech perception. Nonspeech stimuli are generally perceived *continuously* in that listeners' sensitivity to small acoustic differences does not change across a continuum. It was thought that heterogeneous discriminability in speech sounds reflected articulatory discontinuities between different places of articulation and it was therefore presumed that perception reflected these discontinuities. Because the presentation of speech stimuli resulted in discrimination behavior that was strikingly different from that found with nonspeech tokens, many researchers were led to believe that a distinct neural structure existed for the sole purpose of decoding speech into phonetic units and that this was entirely separate from other forms of auditory perception (Liberman and Mattingly, 1985).

The fact that a unique alveolar locus was found at 1800 Hz was used as further evidence that perception was performed on the basis of articulation and not acoustics. It was argued that formant transitions rather directly represented articulatory movements and that the position of the speaker's articulators could be decoded from the acoustic stream. Although it could be argued that the locus itself is an invariant *acoustic* cue, it was shown that when second formant transitions actually did originate at 1800 Hz, the perceived place of articulation was again found to be context dependent (Delattre *et al.*, 1955, see also Sec. 1.3.2.1). This was seen as proof that listeners expect a naturally produced stimulus from a lawfully governed articulatory system—*i.e.*, that the locus target *must* be hidden by the release burst for the stimulus to be perceived consistently.

However, the lack of structural independence between different articulators complicated the relationship between perception and articulation; articulatory invariance was only relatively obvious for the bilabials because of the physical separation between the tongue and lips. Liberman *et al.* (1967) therefore proposed that the *Neuromotor Rules*—or the neural signals that were used in production—held this simple one-to-one relationship with perception. This was viewed as advantageous since the listener could use the same neural mechanisms designed for speech production to decode the acoustic signal in perception. It was suggested that perception was mediated in some way by the neuromotor correlates to articulatory gestures. The neural commands were assumed to be invariant across all contexts despite the observed context dependencies found in acoustics and articulation. As more evidence was obtained regarding the exact nature of neural mechanisms used in speech, the specification of invariant speech units was weakened to merely the gestures *intended* by the speaker. For example, it was known that a simple articulatory movement, such as the raising of the tongue dorsum to make contact with the velum, was the result of several complex neural commands (Liberman and Mattingly, 1985).

The theory of an innate link between production and perception was eventually referred to as the "Motor Theory" of speech perception. Nearey (1992b) refers to this as the *strong gestural* position because of the strong links that are postulated between perception and articulation. However, other theories regarding the relationship between these three domains have been proposed as well. The next section presents *strong auditory* accounts which instead postulate a strong relationship between *acoustics* and perception.

1.4.2 Auditory theories of speech perception and Locus Equations

In apparent contradiction to his colleagues, Delattre (1969)[7] suggested that the characteristic locus for each place of articulation was defined in terms of *perception* and not production in that its frequency value was arrived at by perceptual means. He suggested that although points of second formant convergence could be observed on spectrograms, they had no real relevance unless their perceptual effect could also be accounted for independently of all other potential cues. He pointed out for example, that formant transitions isolated from the burst are often heard incorrectly and that the perception of such burstless stimuli was best understood with reference to an acoustic locus theory. Delattre defined the locus as the "the frequency toward which formant transitions *must* point in order to contribute maximally to the perception of a given place of articulation" (1969, p. 13, original emphasis). Similar to Dorman *et al.* (1977), he also added that coarticulation made the perception of burstless stimuli inviable, and that strong burst cues must therefore compensate. Such cues were defined purely in terms of *auditory* properties and were not at all gestural.

A more extreme position was taken by Cole and Scott (1974b), who suggested that formant transitions were entirely overridden by the spectral properties of the burst and that this acoustic segment contained sufficient invariant information for place of articulation identification. They went on to suggest that the primary role of formant transitions was to provide information regarding the *temporal order* of phonetic segments. As mentioned previously in Sec. 1.3.1, the perceptual experiments performed by Cole and Scott (1974a) included the aspiration segment with the release burst which resulted in an overly generous evaluation of its perceptual significance.

In contrast, Sussman *et al.* (1998) presented a summary of evidence that second formant transitions might serve as an invariant cue to the perception of stop consonant place of articulation when the vocalic context is taken into account. It was originally found by Lindblom (1963) that the onset frequency of F_2 was strongly correlated with the F_2 of the following vowel and that the slope and y-intercept of the regression fit was specific to the place of articulation of the consonant (see also Sussman, 1991; Sussman *et al.*, 1991). This was a much weaker claim than the *Locus theory* proposed earlier by Delattre *et al.* (1955, see also Sec. 1.3.2.1) in that the onset and vowel second formant frequencies were assumed to be merely *correlated* and that formant transitions could not necessarily be extrapolated to a single unique locus. These correlations and their regression fits have been referred to as *locus equations*.

[7] who coauthored many papers with Liberman (Cooper *et al.*, 1952; Liberman *et al.*, 1952, 1954; Delattre *et al.*, 1955; Liberman *et al.*, 1958; Harris *et al.*, 1958)

Sussman *et al.* (1998) gave a thorough overview of locus equations as a partly invariant cue to the perception of stop consonant place of articulation. Not only did they suggest that perception of stop place was defined in purely acoustic terms, they even proposed a neurological mechanism whereby the human auditory system could take advantage of the observed correlations. They did so by drawing parallels between this language specific ability and perceptual mechanisms that have been studied in specific mammalian and avian species. They also showed that this acoustic patterning was not merely a byproduct of human articulatory physiology, but that speakers actually exploited this acoustic correlation by specifically targeting the locus equation relations in articulation. It has also been shown that they also provide good correlates to perception in synthetic speech (Shammass, 1985; Fruchter and Sussman, 1997).

Unfortunately, the proposal for a specific neurological mechanism that underlies this behavior detracted from the main point of their research which was to establish locus equations as an important acoustic correlate. It also drew tremendous criticism (see Sussman *et al.*, 1998). Specifically, Sussman *et al.* compared stop consonant perception in humans to echo-location in bats—which relies on the detection of the magnitude of Doppler shift in very short wavelength signals—and to interaural time difference detection in barn owls. For these and many other species, simple combination-sensitive neural structures have already been identified.

They also proposed that human listeners' ability to distinguish stop consonant place of articulation may have evolved from such mechanisms. However, there are many differences between the perceptual abilities of bats and barn owls and speech perception in humans. Echo location and time difference detection are necessarily perceived continuously and must also be unambiguous and nonoverlapping. In contrast, it is known that stop consonant place perception is very nearly categorical (Liberman *et al.*, 1957; Eimas, 1963) and that the distributions of F_2 onset and vowel steady state frequencies show a large degree of between-category overlap.

With regards to the problem of overlapping category distributions, Sussman *et al.* (1998) argued that locus equations represent only part of the full set of acoustic correlates to place of articulation perception and that other cues, such as the burst (Sec. 1.3.1) and F_3 formant frequency (Sec. 1.3.2.3), must be taken into account to increase the perceptual separability between phonemes (*cf.*, Lindblom, 1990, see also Sec. 1.3.2.3). In fact, Nearey and Shammass (1987) found similar locus equation relationships for the third formant as well, which improved the identification of prevocalic stop consonants in an automatic classification experiment.

Nevertheless, Sussman *et al.* (1998) demonstrated the robustness of locus equation regression fits across speakers and across different speaking rates. However, quite a bit of additional variation was introduced which further exacerbated the overlap between categories. While, Sussman *et al.* reported modest correct classification rates using the locus equation relations, they added that when the *slopes* and *y-intercepts* were used in discrimination themselves, correct classification rose to a full 100%. However, since full locus equations can not be determined from a single utterance, they are not useful in discrimination. It also implied that some kind of speaker normalization must take place in order to perceive place distinctions.

Another criticism to the perceptual relevance of locus equations was that the correlations between onset and vowel F_2 may simply be a consequence of articulation and that speech is constrained by coarticulatory effects between the vowel and consonant. There is no reason why the human perceptual system could not take advantage of this articulatory covariation. However, because Sussman *et al.* (1998) presented a purely perceptual/neurological motivation for this phenomenon, the point was made by Fowler (1994) that this line of reasoning is completely superfluous if there already exists a well-justified articulatory explanation. Nevertheless, it has been shown that speakers fitted with bite blocks to perturb the normal functioning of articulation produce the same kind of lawful variability as is found in normal speech (Sussman *et al.*, 1995). However, additional influences, such as the preceding vowel context (Öhman, 1966) have also not been accounted for in Sussman *et al.*'s research.

1.5 Gross spectral shape cues

Based on the acoustic theory of speech production, Fant (1970) suggested that place of articulation could be derivable from the burst and the first 20 ms of formant transitions. Several authors have taken this view and have subsequently proposed a set of cues based on the global spectral properties of the first few tens of milliseconds following the onset of the release burst.

Section 1.5.1 describes theories that have been proposed along these lines. However, several researchers, unsatisfied with the apparent paucity of information provided by these brief segments of time, have instead proposed dynamic spectral shape theories which are presented in Sec. 1.5.2.

Unlike detailed theories of speech perception, global spectral shape theories do not treat burst, aspiration or formant transition segments differentially. They also presuppose a direct and simple relationship between articulation and acoustics as well as between acoustics and perception. Nearey (1992b) refers to this class of hypotheses as *double strong* theories because of the assumed direct links between all three domains.

1.5.1 Onset spectral shape

Stevens and Blumstein (1978) suggested that the first 10–20 ms of the spectrum sampled at the onset of oral release was most likely to contain context invariant information regarding place of articulation. In the case of absent, or very short periods of frication, the burst and vocalic formant transitions were thought to form a single integrated acoustic property that was invariant within places of articulation. Bilabial releases were characterized by low-frequency formant energy resulting in a downward tilt in spectral shape. Similarly, alveolar articulations were characterized by higher frequency formants which resulted in a rising spectral tilt (*cf.*, Sec. 1.3.1). Because the second and third formants were relatively close for velars, their proximity resulted in a relatively pronounced peak of energy in the mid-frequency range (*i.e.*, between 1–3 kHz), whereas bilabial and alveolar spectra were more diffuse. Hence it was thought that place of articulation could be discriminated on the basis of two simple parameters: spectral slope and compactness.

Stimuli were synthesized to test this hypothesis. When F_2 and F_3 were less than 780 Hz apart, subjects responded mostly with [g]. Otherwise, the spectral balance between low and high-frequency peaks determined the identification of either [b] or [d]. However, it is clear that the use of spectral peaks in the bursts of synthesized stimuli ensured that detailed and gross spectral cues were correlated, and it is not clear how such experiments confirm the relative importance of global spectral shape *vs.* burst peak frequency information.

A similar problem is found in experiments performed by Blumstein and Stevens (1979) who used templates to categorize the place of articulation of syllable-initial stops. The templates match prominent spectral peaks found in the spectra of release onsets. For example, if any two such peaks were found to be within 500 Hz of each other, the corresponding spectrum was labelled as velar. Otherwise, multiple peaks above 2200 Hz were labelled as alveolar and a peak between 2400–3600 Hz were labelled as bilabial.

Blumstein and Stevens (1980) suggested that the deletion of release bursts does not change the global spectral properties of the onset of the remaining signal. It is clear however, that the removal of such cues degrades correct identification performance significantly (Sec. 1.3.4.3). Despite this, Blumstein and Stevens claimed that although the burst did provide some additional place information, it was not "essential", in that its deletion did not result in completely random identification behavior.

Although the above cited authors presented some evidence for an invariant acoustic cue to place of articulation in the onset of stop release bursts, Suomi (1985) pointed out that many of the gross spectral templates used in these studies ignored vowel specific coarticulatory effects in the burst and therefore hid the contextual variability that is indeed truly there. The question was whether this additional information actually improved the discriminability of different places of articulation or not. Suomi found that automatic classification using a multivariate mixture distribution on the basis of vowel context performed better than a single context-invariant category center for each stop consonant.

1.5.2 Temporally dynamic spectral shape

Prior to the experiments described in the previous section, Stevens (1975) had originally suggested that consonants were characterized primarily by rapid spectral changes within the first 20–30 ms of the syllable onset. These spectral changes were related to second formant transitions: labials show a frequency rise in gross spectral peaks, while alveolars show a frequency drop (*cf.*, Sec. 1.3.2.1). Velars on the other hand, show a spectrally diverging pattern similar to what had been described before (Potter *et al.*, 1947; Fischer-Jørgensen, 1954, see also Sec. 1.3.2.3). Similar to Stevens and Blumstein (1978) and Blumstein and Stevens (1979, 1980), however, global spectral shape patterns were primarily defined in terms of the behavior of detailed spectral features, such as formant transitions, which entailed that it was impossible to distinguish the effects of formants and burst peaks from gross spectral properties.

Using a spectral representation motivated by the physiological properties of the human auditory system, Searle, Jacobson, and Kimberley (1980) were able to classify place of articulation for a test set of prevocalic voiceless stops in continuous speech with 80% accuracy. The spectral analysis calculated the spectro-temporal envelope of the acoustic signals and evaluated consonant identity on the basis of a few discriminant features, such as the abruptness of the burst onset, VOT, and the rate of formant transitions, in addition to the location of the burst peak. Similarly, Kewley-Port (1983) and Kewley-Port and Luce (1984) characterized place of articulation in terms of time-varying relational acoustic features that were defined largely in terms of the global spectral tilt and compactness at stimulus onset as well as the onset time of low-frequency energy associated with F_1-cutback. Kewley-Port *et al.* (1983) further synthesized two sets CV syllables conforming either to the static spectral properties described by Stevens and Blumstein (1978) or to the dynamic spectral properties put forth by Kewley-Port (1983). Listeners identified temporally dynamic stimuli significantly better. They claimed, therefore, that acoustic cues for place of articulation are located in the first 20–40 ms, which primarily takes into account differences in VOT for the three places of articulation as well as the continuation of compactness over longer durations for velars. However, it was found that [b] and [d] could otherwise be distinguished on the bases of the first 20 ms alone.

Based on the measurement of natural stimuli, as well as the results of synthesis experiments, Lahiri *et al.* (1984) concluded that it was not the absolute global spectral shape of the first few tens of milliseconds, but rather the *relative* change of this property over time that was best able to discriminate places of articulation. For example, it was hypothesized that there was a smaller change in high-frequency energy between the burst onset and the onset of voicing for alveolars than for bilabials.

Two automatic classification experiments have exploited spectro-temporal information in prevocalic stops. In addition, they also compared the relative usefulness of static *vs.* dynamic spectral features. Using the mean, skewness, and kurtosis of the normalized power spectrum from a single 20 ms analysis frame centered at the onset of the burst, Forrest *et al.* (1988) found that a linear discriminant analysis was able to classify the place of articulation of voiceless stops produced by female speakers with 87.1% accuracy when trained on tokens produced by male speakers. This value increased to 91.2% when an additional 20 ms analysis frame centered 10 ms following the onset of the burst was included, and then to 93.5% when another two analysis frames were added. Using both formant frequencies and amplitudes, as well as discrete cosine transform coefficients (DCTCs), Nossair and Zahorian (1991) found that a quadratic discriminant analysis from static burst spectra was able to identify correct place of articulation of either voiced or voiceless stops from an independent test set with 82% and 73% accuracy for the DCTCs and formants respectively. However, when DCTCs were measured at several frames with a total effective duration of 60 ms, this value increased to 95%. Both experiments show therefore, that dynamic spectral features offer significant benefits to the automatic classification of prevocalic stops.

1.6 Explicit comparisons of detailed *vs.* gross spectral properties

In order to make a more direct comparison between the relative perceptual importance of either global spectral shape properties or detailed acoustic cues, several experiments were designed using a conflicting-cues paradigm in which gross and detailed properties were appropriate for different places of articulation. Blumstein, Isaacs, and Mertus (1982) manipulated formant frequencies and spectral tilt independently in stimuli beginning with either [b] or [d] before the vowels [i], [ɑ], and [u]. They found that there was a significant drop in responses appropriate for detailed cues when spectral tilt conflicted with formant patterns. Overall however, it was found that the majority of responses were appropriate for the place of articulation specified by the formant frequencies. They suggested that gross spectral shape may play a role only in relation to spectral changes over time. Walley and Carrell (1983) independently manipulated both formant transitions and onset spectral shape in synthetic burstless stimuli within two vowel contexts [u] and [ɑ] by manipulating synthesis formant amplitudes. Subjects classified conflicting cue stimuli primarily on the basis of formant transitions. It was additionally found that this was true for both child and adult subjects. Similar to the findings of Blumstein *et al.*, however, responses appropriate to formant transitions did decrease in conflicting-cue stimuli. Similarly, Lindholm *et al.* (1988) found that normal-hearing speakers relied more on formant transition information than spectral shape in place judgments (Sec. 1.1). In another conflicting-cue paradigm in which naturally produced speech was manipulated parametrically, Dorman and Loizou (1996) found that listeners were more likely to categorize stimuli on the basis of formant transition information rather than relative changes in spectral tilt as suggested by Lahiri *et al.* (1984).

Another approach in comparing gross and detailed spectral features, has been to model listeners' responses using explicit detailed or gross cue hypotheses. For example, Krull (1990) compared place confusions in subjects' identifications to Euclidean distances between place category centers in either gross or detailed feature spaces. She found that relative distances based on onset formant frequencies of F_2–F_4, when combined with burst duration, were better able to predict listeners' confusions than gross spectral shape, represented by filterband spectra. Similarly, Smits *et al.* (1996b) found that detailed cues, such as formant transitions, gave a better fit to listeners' responses to cross-spliced, naturally produced stimuli in which the burst and vocalic segments were concatenated from different places of articulation.

However, in contrast to all of the results cited above, Nossair and Zahorian (1991) found that global spectral shape properties in the form cepstral coefficients provided better stop classification performance than formant frequencies in an automatic classification experiment (see also Sec. 1.5.2).

1.7 Distinguishing between gross and detailed cues

The research presented in this thesis approaches the controversy between gross and detailed spectral cues in the perception of prevocalic stop consonants using a completely different experimental paradigm. As it has been assumed that detailed spectral features require a greater spectral resolution than gross features such as spectral tilt and compactness, differences between the two accounts of perception might be more apparent for stimuli that have been spectrally reduced in such a way as to effectively remove detailed properties such as burst peak and formant frequencies. This can be compared to a deleted cue experiment in which listeners are asked to categorize speech stimuli in which some potentially important phonetic information has been removed—*i.e.*, in this case detailed spectral features have been distorted to the point where such properties are effectively eliminated.

Before the opposition between gross and detailed acoustic features is addressed directly, Chap. 2 explores the importance of dynamic spectral information in short, gated stop bursts. This is related to the question of whether spectral information that evolves over time is perceptually important in stop place-of-articulation identification as proposed by researchers such as Kewley-Port (1983), or whether a static representation is sufficient as suggested by Stevens and Blumstein (1978) and Blumstein and Stevens (1979, 1980).

Chapter 3 first explores whether detailed spectral features, such as burst peak and formant frequencies, are potentially recoverable from spectrally reduced stimuli. It is subsequently shown that while some such information is still present in these stimuli, a gross spectral cue account is better able to model listeners' categorizations. Chapter 4 then addresses the question of whether spectrally reduced formant transitions alone are able to convey sufficient place-of-articulation information. Despite the fact that no formant frequency information can be obtained from such distorted stimuli, it is shown that listeners are still able to identify the place of articulation of prevocalic stops significantly better than chance, suggesting that subjects in these experiments attend primarily to spectral shape information in spectrally reduced signals. This conclusion is further supported by modeling results which show that spectral shape features are better able to predict listeners' responses to such stimuli.

Chapter 2

Temporal information in gated stop consonants

2.1 Introduction

The goal of this chapter is to assess the importance of *dynamic* spectral information in short gated stop bursts. It has been suggested by Fant (1970) that the configuration of the articulators within the vocal tract at the moment of stop-consonant oral release produces an acoustic output that should be sufficient for place of articulation discrimination in these speech sounds. This hypothesis has led several researchers conclude that most of the phonetically relevant acoustic information is contained within the first 10–20 ms of the speech signal immediately following oral release—possibly including the subsequent vocalic portion (Stevens and Blumstein, 1978; Blumstein and Stevens, 1980). However, this position has been challenged numerous times on the grounds that spectrally dynamic information is additionally necessary for correct identification by listeners (*e.g.*, Blumstein, Isaacs, and Mertus, 1982; Kewley-Port, 1983; Kewley-Port, Pisoni, and Studdert-Kennedy, 1983; Kewley-Port and Luce, 1984; Walley and Carrell, 1983; Lahiri, Gewirth, and Blumstein, 1984). For example, it has been shown that voice onset time (VOT) is at least partly contrastive and may help to distinguish velars from other places of articulation (Kewley-Port, 1982; Lisker and Abramson, 1964, 1967). Similarly, it is also believed that formant transitions are important cues to stop perception as well (Cooper *et al.*, 1952; Liberman *et al.*, 1954; Sussman *et al.*, 1998).

Spectrally dynamic models of stop consonant perception are typically based on much longer stimulus durations: either a fixed length (*e.g.*, 40 ms, Kewley-Port, 1983) or a variable length based on the duration of the burst or VOT (*e.g.*, Lahiri *et al.*, 1984). Nevertheless, it is known that listeners are quite capable of identifying gated stops of 10–20 ms in duration (Tekieli and Cullinan, 1979) and that classification based on the burst alone is statistically better than chance (*e.g.*, Fischer-Jørgensen, 1972; Smits, ten Bosch, and Collier, 1996a). The two views are to a certain extent quite compatible: it is possible that the perception of short, gated prevocalic consonants is determined by a static representation of the acoustic stimulus, while longer signals introduce perceptually relevant dynamic information such as VOT and formant transitions. Clearly, listener identification of stop place is significantly better with much longer stimuli (Nossair and Zahorian, 1991), and the fact that spectrally dynamic information is perceptually relevant at longer durations has been demonstrated (Kewley-Port *et al.*, 1983).

While temporally dynamic information is often associated with acoustic cues such as formants or VOT, it is not known if such information is present at smaller time scales in the order of 10–20 ms. Experiments in speech perception and automatic classification performed by Stevens and Blumstein (1978) and Blumstein and Stevens (1979, 1980) used power spectra from the initial portion of such stimuli under the assumption that dynamic spectral features within these timespans are not used to discriminate place of articulation by human listeners. However, it has also been shown that some temporal information may be useful in even very short stimuli such as isolated bursts. For example, Cole and Scott (1974b) suggested that the *waveform envelope* of the release burst could be used to discriminate [p] from [k]. This may be related to the finding that velar bursts often exhibit multiple releases (Fischer-Jørgensen, 1954; Dorman, Studdert-Kennedy, and Raphael, 1977). Temporal information such as the waveform envelope are ignored by analyses that focus on the spectra of release onsets alone; these theories assume that the release burst is a stationary acoustic cue to place of articulation. However, a static representation of the first few milliseconds following oral release may not be adequate to predict listeners' perception of even short, gated stop bursts.

In order to test the hypothesis that short, gated bursts contain only static spectral information, two experiments were designed. In the first experiment, naturally produced stop consonants were classified via a linear discriminant analysis (LDA) using either a temporally static or spectrally dynamic representation of the first few tens of milliseconds following the burst onset. This simple test was designed to show whether temporal information is *potentially* useful in perception. Following this, a perceptual experiment was designed to evaluate the perception of short, *temporally distorted* gated stop bursts by comparing the correct identification rates of these stimuli to undistorted tokens. The spectrally processed stimuli in this test were altered by distorting the waveform envelope while simultaneously preserving the long-term power spectrum, or static spectral characteristics; hence temporal information was effectively removed. Under the assumption that no spectrally dynamic information is used by listeners in the categorization of these short stimuli, we would expect that this transformation will have little or no effect on overall correct identification performance. Conversely, a change in perception would indicate that some dynamic information is being exploited by listeners to determine the phonetic identity of these stimuli.

It should be pointed out however, that while these manipulations may produce a perceptible change in the *quality* of the stimuli, this fact alone does not necessarily entail that psychologically relevant *phonetic* information has been disturbed in any way. The goal of this work is to reduce gated bursts to only their discriminant acoustic features by discarding or distorting phonetically irrelevant information. A null result—*i.e.*, one which shows that listeners are unaffected by changes in temporal information—would indicate that a static representation of gated stop bursts is completely adequate for the perception of their phonetic identity.

Therefore, the goal of this chapter is to compare perception performance between unmodified and temporally distorted stimuli in order to evaluate the relative importance of dynamic spectral information in gated stop bursts. Before this problem is addressed directly, an automatic classification experiment is performed to evaluate the potential discriminability of static *vs.* dynamic featural representations. This is presented in Sec. 2.2, while the perceptual experiment is described in Sec. 2.3. Based on these experiments, it is hypothesized that the source of much of the temporal information in short gated stimuli can be attributed to the relative onset of voicing or VOT. An alternative analysis based on this assumption is presented in Sec. 2.4. Finally, Sec. 2.5 discusses any conclusions that can be drawn from this work.

2.2 Experiment 1: Automatic classification

The purpose of this experiment is to evaluate the relative performance of both static and dynamic spectral representations of gated stops in an automatic classification task using a database of naturally produced tokens. A linear discriminant analysis was used to classify these tokens by place of articulation using one of two sets of acoustic features which differed in the number of frames analyzed. The first set of features is spectrally *static* in that Mel-frequency cepstral coefficients (MFCCs) were measured from the full length of the gated stimuli. In the second set, MFCCs from two frames were measured. Using a relatively unbiased estimate of their correct classification performance, it is then possible to evaluate whether temporally dynamic information is *potentially* available to listeners from such stimuli. If this is indeed the case, then it is possible that human listeners exploit such information in perception.

2.2.1 Speech sample

The stimuli used in this experiment were recorded from 14 native speakers of Western Canadian English (five males and nine females) consisting entirely of CVC syllables. The initial consonant was drawn from [p], [t], [k], [b], [d], and [g], while the final consonant was one of [k], [ʤ], or [l]. The set of medial vowels consisted of the full inventory of 14 stressed vowels found in this dialect of English: [i], [ɪ], [e], [ɛ], [æ], [ʌ], [ɒ], [o], [ʊ], [u], [ɚ], [aɪ], [aʊ], and [ɔɪ]. An exhaustive set of all combinations of these phonemes would result in 3528 stimuli. However, several tokens had to be removed because of recording or other difficulties leaving only 3079.

The recordings were lowpass filtered at 7.8 kHz and digitized at 16 kHz using 12 bit quantization. These particular stimuli were originally used in an automatic classification project by Nearey (1992c).

The onsets of these stimuli were located interactively using a waveform and spectrogram display and the preceding silence was removed. Some stimuli (in particular voiced bilabials) contained a prerelease voicebar; in these cases the prevoicing was also removed by cutting the stimuli at a zero crossing just before the onset of the release burst. The stimuli were subsequently gated (*i.e.*, truncated) at 11 different durations ranging from 13.75–63.75 ms in 5 ms steps. The final 5 ms of each gated stimulus was tapered by the second half of a 10 ms long Hamming window to reduce spectral distortion that would otherwise result in transients or "pops" at the point of truncation. Figure 2.1 on the following page illustrates the gating procedure. The upper waveform represents the original utterance [dɒk] as spoken by a female speaker. The dashed lines in the top graph indicate the points of truncation—the first was selected by visual inspection and, in this example, the second is located exactly 23.75 ms following. The bottom graph gives the resulting gated waveform after weighting with the half-Hamming window.

Each gated stimulus was analyzed in either one or two frames. In the one-frame condition, spectral features were measured from the full length of the gated waveform. In the two-frame condition, the stimuli were divided into two frames of equal length

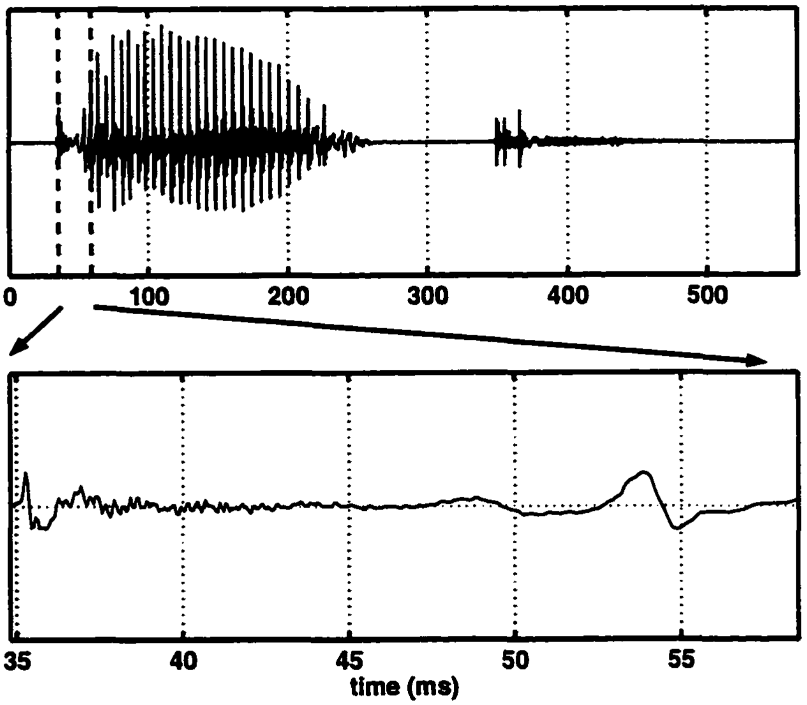


Figure 2.1: Illustration of gating procedure. The top graph gives the waveform of the syllable [dɒk] as produced by a female speaker while the bottom graph gives the waveform from the first 23.75 ms following the burst onset. The dotted lines in the top graph show the limits of the gated stimulus.

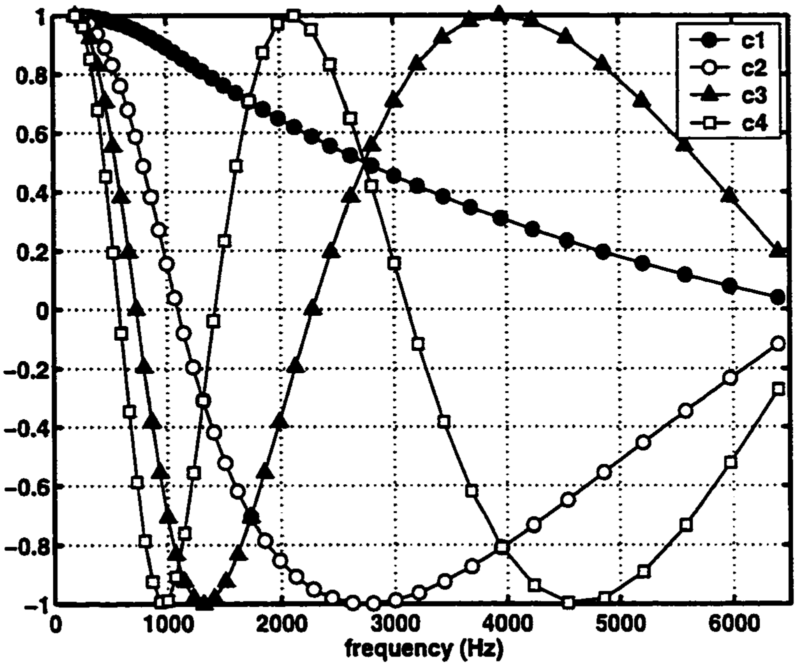


Figure 2.2: First four discrete cosine basis functions of MFCCs. The markers on the function curves indicate the center frequencies of the 40 triangular filters used to transform the spectrum into the Mel-frequency scale.

and each frame was analyzed separately. The two frames in this condition overlapped by 5 ms and the last 5 ms of the first frame was tapered by a half-Hamming window as was the first 5 ms of the second frame.

For each analysis frame, the signal was preemphasized by 6 dB/octave and analyzed by a bank of 40 triangular filters spaced approximately equally apart in the Mel-frequency scale with center frequencies ranging from 200–6400 Hz. The first 13 discrete cosine coefficients (DCCs) from the magnitude spectra of this filterbank analysis are used in the subsequent LDA. The implementation used for the extraction of these parameters was written by Slaney (1994) for MATLAB.

The basis functions for the first four MFCCs are given in Fig. 2.2. The first basis function measures the overall spectral slope, while the second gives the relative difference in energy between the mid-frequency and low-frequency regions. Together, they are expected to provide a rough estimate of spectral tilt and compactness or peakedness of the mid-frequency region (Stevens and Blumstein, 1978; Blumstein and Stevens, 1979, 1980). Higher coefficient basis functions allow for additional spectral complexity.

2.2.2 Discriminant analysis

The stimuli described in the previous section were automatically classified using a leave-one-speaker-out cross-validiation—*i.e.*, the tokens produced by each individual

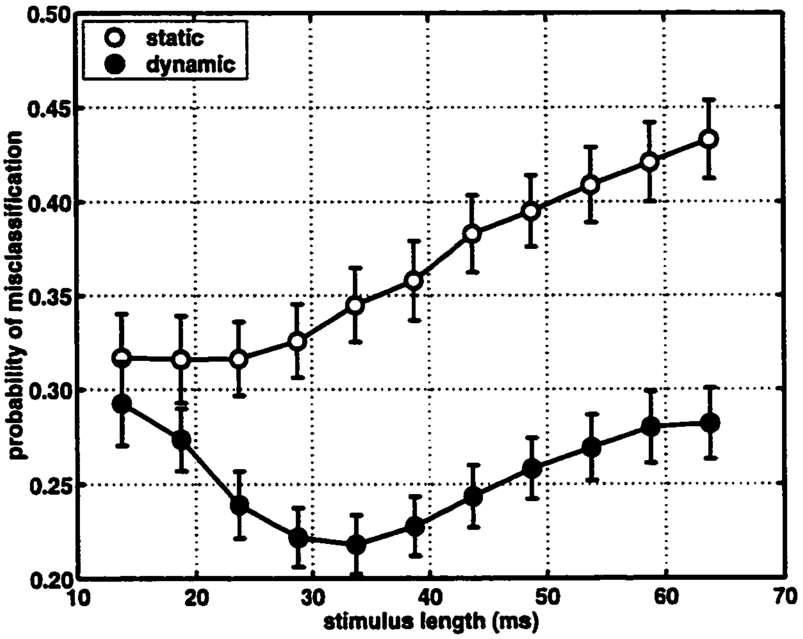


Figure 2.3: Comparison of static *vs.* dynamic parameters in an automatic stop place classification task. The graph shows probability of misclassification (pmc) for place of articulation identification. The error bars represent ±1 standard error based on the between-speaker standard deviation of the mean pmc.

speaker were categorized on the basis of an LDA estimated from the tokens from the remaining 13 speakers.

2.2.3 Results

Figure 2.3 shows the probability of place of articulation misclassification for both the static and dynamic feature sets. The dynamic representation performs consistently better. However, the errorbars, which indicate ±1 standard error based on the probability of misclassification for each speaker, show that this difference is not significant for the shortest gating duration and only marginally so at 18.75 ms. The difference between the static and dynamic parameter sets is highly significant at 23.75 ms. It can be said then that some spectrally dynamic information can be used to discriminate place of articulation in gated stop bursts except at the very shortest gating durations used in this experiment. Such information is therefore potentially available to human listeners as well. The next section evaluates the hypothesis that spectrally dynamic information might be used in the perception of short, gated stop bursts.

2.3 Experiment 2: Perception experiment

Although dynamic spectral information can be used by an automatic classifier to improve discrimination of place of articulation, the question of whether humans use this kind of information in speech perception will now been addressed. In order to evaluate the importance of such cues, an experiment was designed in which temporal information within short gated bursts is distorted while static spectral information is preserved. If human listeners ignore the dynamic spectral properties of such stimuli then they should be relatively insensitive to these temporal changes and the distortion of the waveform envelope will have no effect on their ability to identify the original place of articulation. Therefore, listener perception of unprocessed-gated stops is compared with the classification of similar stimuli with altered dynamic spectral detail.

2.3.1 Stimuli

The stimuli used in this experiment are similar to those described in Sec. 2.2.1. They were recorded from 12 speakers of Western Canadian English (five males and seven females) and are all of the form CVC with the additional constraint that the final syllable is always [k]. In order to further keep the total number of stimuli small, the medial vowels used in this experiment were restricted to the set [e] and [o], which represent the extremes of F_2 in this dialect of English (Nearey and Assmann, 1986), as well as [æ] and [ɒ] which were included as intermediate vowels in the total range of F_2. Based on the results in the previous section, stimuli were gated at two durations: 13.75 and 23.75 ms with the tails tapered by 5 ms long half-Hamming windows. In total there were 288 original unprocessed stimuli for each of the 13.75 and 23.75 ms gating conditions. Because the tails of these stimuli have much lower amplitudes, these conditions will henceforth be referred to as 10 ms and 20 ms respectively.

2.3.2 Signal processing

The stimuli described in the previous section represent the *unprocessed-gated* condition. A second set of stimuli was generated having the same long-term power spectrum but different waveform envelope. This was accomplished by determining the minimum-phase component of the minimum-phase/allpass decomposition for each of the gated stimuli (Oppenheim and Schafer, 1989). A nonminimal-phase signal can be represented as

$$x[n] = x_{\mathrm{min}}[n] * x_{\mathrm{ap}}[n] \tag{2.1}$$

where $x_{\mathrm{min}}[n]$ is a minimum-phase signal, $*$ is the convolution operator, and $x_{\mathrm{ap}}[n]$ is the impulse response of an allpass filter having $|X_{\mathrm{ap}}(e^{j\omega})| = 1$ for all ω—*i.e.*, the magnitude response is maximally flat. The minimum-phase component, $x_{\mathrm{min}}[n]$ is characterized by ***minimum energy delay*** meaning that the power of the signal is greater at the onset. This is due to the fact that the Fourier phase of the signal is monotonic decreasing and all of the zeros of the signal's z-transform are found inside the unit circle on the z-plane. The allpass filter $x_{\mathrm{ap}}[n]$ inverts some of these zeros resulting in the original nonminimum-phase signal. Intuitively, the minimum-phase component consists of the same spectral components found in the original signal, but with most of the energy shifted backwards in time towards the beginning. Because the magnitude response of

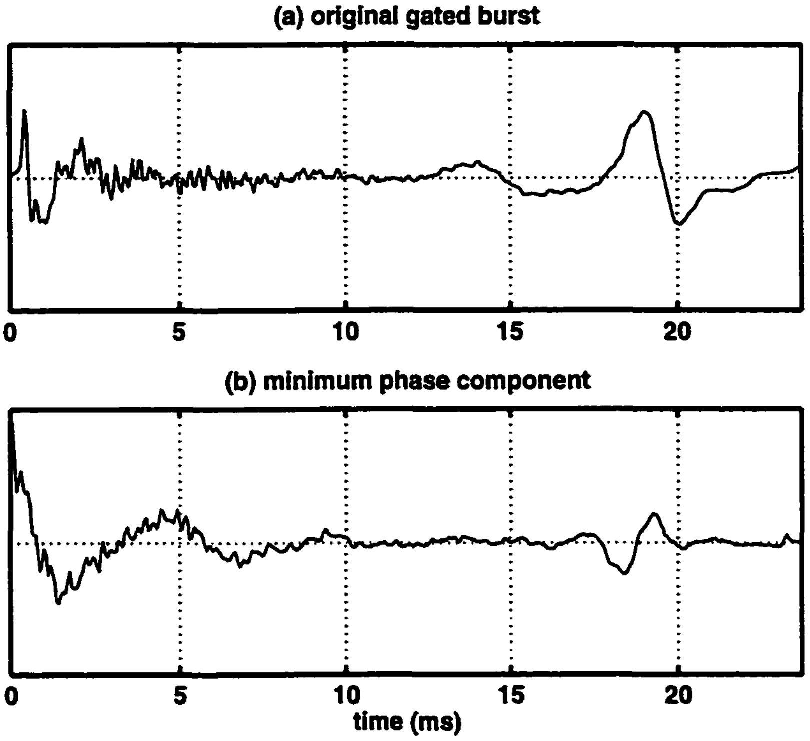


Figure 2.4: Original and minimum-phase component waveforms. The figure compares the original stimulus waveform (a) to that of the minimum-phase component of minimum-phase/allpass decomposition (b). The token is the same as that found in Fig. 2.1 on p. 34. The top stimulus was most often heard as [d] (5/6 times), while the bottom stimulus was most often heard as [b] (also 5/6 times—see Sec. 2.3.5).

the allpass component is maximally flat, the minimum phase component must have the same Fourier magnitude as the original signal—*i.e.*, $|X(e^{j\omega})| = |X_{\min}(e^{j\omega})|$ for all ω.

The graph in Fig. 2.4(a) on the current page again shows the gated waveform found in Fig. 2.1 on p. 34, while Fig. 2.4(b) shows its minimum-phase component. Note that the energy at the onset of the minimum-phase component is amplified while the remainder of the stimulus is relatively attenuated due to the minimum energy delay property described above.

Figure 2.5 on the next page shows that the magnitude spectra from the unprocessed example and its minimum-phase component are nearly identical. The difference between the two spectra can be seen in Fig. 2.6(a) on p. 40 which shows the magnitude response of the filter derived from the allpass component. While we expect a completely flat magnitude response, in actuality this is not the case because of the 16 bit quantization used to digitize the stimuli. However, when compared with Fig. 2.5 on the next page, it can be seen that the largest deviations from 0 dB magnitude response correspond to zeros in the original spectrum. When the minimum-phase component is stored in double precision floating point format, such distortion is in the order of $10e^{-12}$dB.

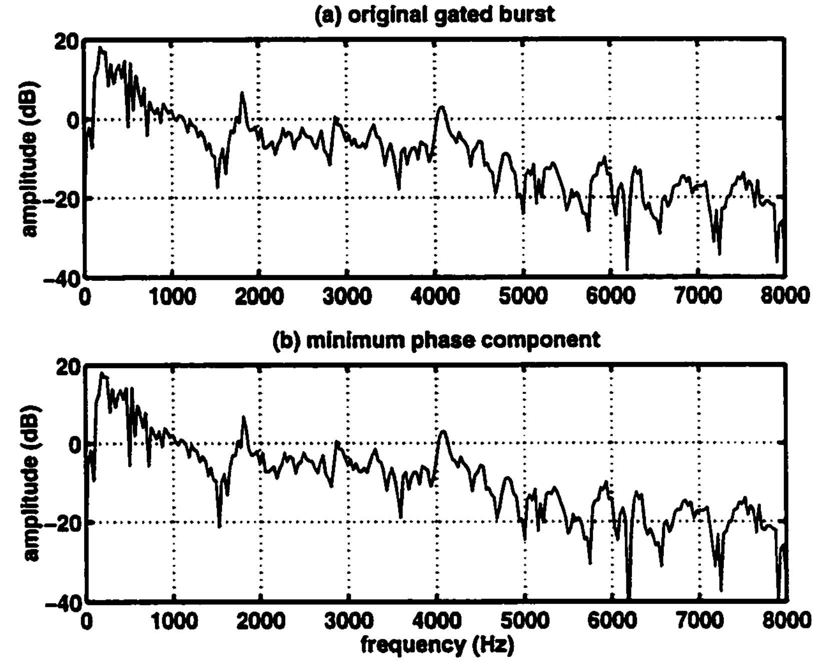


Figure 2.5: Original and minimum-phase component spectra. The figure compares of power spectra from the original stimulus (a) and the minimum phase component (b) of the waveforms illustrated in Fig. 2.4 on the preceding page.

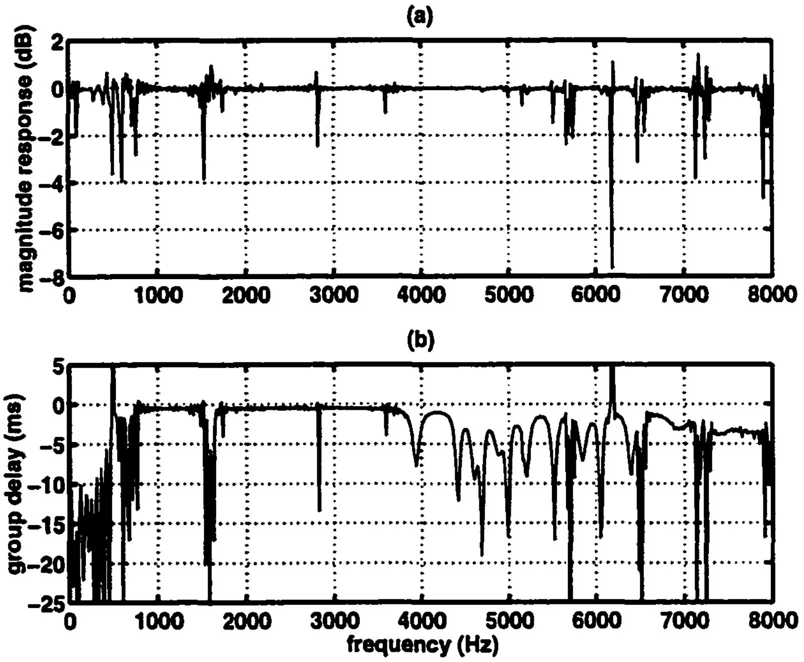


Figure 2.6: Magnitude and group delay response of allpass component of minimum-phase/allpass decomposition of the gated waveform shown in Fig. 2.1 on p. 34.

Instead of providing the phase response of the allpass component, the group delay is given in Fig. 2.6(b). This shows the overall shift in spectral energy in milliseconds. The figure shows that most of the delay is negative (with a couple of positive components corresponding to spectral zeros and are largely due to quantization error) indicating that much of the spectral energy is shifted backwards in time toward the onset of the signal.

The minimum-phase component can be estimated by homomorphic filtering (Oppenheim and Schafer, 1989): the positive indexed coefficients of the signal's real cepstral transform are zeroed and the minimum-phase signal is produced via the inverse cepstral transform. In order to avoid aliasing effects, it is necessary to pad the original signal, since the effective duration of the impulse response of the allpass component can be extremely long. Homomorphic filtering was performed using the RCEPS function in MATLAB.

Originally, several other techniques were employed to generate stimuli for the temporally distorted condition. However, none of these alternative methods proved satisfactory. For example, unconstrained manipulation of the Fourier phase results in temporal aliasing. Another procedure that was attempted used simple allpass filtering to randomly distort the the Fourier phase. However, it is known that humans are relatively insensitive to small changes in phase, and the complexity of the allpass filter that was required was therefore so great that it resulted in a dramatic lengthening of the resulting stimuli making them easily distinguishable from the original tokens.

The use of the minimum-phase component was justified on the basis of a production model of stop production. It has been suggested that the spectral prominences found in release bursts are primarily determined by the size and shape of the oral cavity anterior to the point of constriction in the vocal tract (Fant, 1970; Stevens, 1993). Although it is clear that spectral zeros are also present, this simplifying assumption entails that burst spectra can be well approximated by an allpole model. If we also assume that release bursts are stationary, then an appropriate *causal allpole model* must necessarily also be minimum-phase to ensure stability. If it is the case that all of these assumptions regarding the production of release bursts are correct, then the unprocessed gated stimuli must already be minimum-phase; this would therefore entail that the waveform envelopes of short, gated stop consonants can be determined entirely by their power spectra and that there is *no temporal information* that is not completely redundant. If some of these assumptions are false (as they almost surely are), then the minimum-phase component will differ in at least some temporal information. This is indeed likely to be the case because of the presence of burst frication following the transient release. Frication energy is not necessarily a stationary process and likely falls under the influence of articulatory movement resulting in dynamic changes in resonance frequency.

If it happens that such dynamic spectral information is important in the perception of these stimuli, then modification of the waveform envelope in such a manner will have a deleterious effect on correct identification performance by human listeners. However, if the experiment produces a null result—*i.e.*, one in which no observable effect on perception is produced—this would indicate one of two things: that listeners cannot exploit temporal information in gated stop consonants because there is no such discriminant information available, or that listeners simply do not use such information in phonetic categorization even though this information is potentially present.

An intermediate possibility is that the original stimuli are very nearly minimum-phase already, in which case listeners' identifications will remain unchanged. It may nevertheless be the case that listeners do exploit temporal information, but that the similarities between the unprocessed-gated and minimum-phase stimuli are so great that phonetically relevant dynamic detail is effectively preserved. If this is the case, however, then we have at least discovered a single *static* representation that is able to capture these phonetically relevant details which is, in some sense, a transformation of the auditory spectro-temporal properties attended to by listeners in speech perception. Such a result would provide a basis for further research in the cognitive representation of stop place of articulation. As we shall see in Sec. 2.3.5, this includes modeling listeners' responses based on an explicit theory of perception.

Section 2.2 established that some spectrally dynamic information is indeed present in these stimuli at durations of 23.75 ms. The following experiment attempts to evaluate the usefulness of such information in perception as well as whether some information exists at shorter durations that were not captured by the LDA.

2.3.3 Subjects

The subjects used in this experiment were six graduate and undergraduate students of Linguistics who were native speakers of Western Canadian English and who reported no hearing impairment. They were paid for their participation.

2.3.4 Procedure

Each subject participated in two sessions each consisting of either the 10 ms or the 20 ms stimuli. Unprocessed-gated and minimum-phase stimuli were randomly mixed in each session. Each stimulus was presented over a loudspeaker in a sound-attenuated room. With each presentation, subjects were required to respond with either "b", "d", "g", "p", "t", or "k" by clicking on the appropriate button on a computer screen. Each token was identified in this manner once by each subject resulting in a total of 576 classifications per subject for each of the 10 ms and 20 ms conditions.

Each session was preceded by a test session which was intended to familiarize the subjects with the types of stimuli that were to be presented. No feedback was given during this session.

Only five of the six subjects actually participated in both the 10 ms and 20 ms conditions.

2.3.5 Results

Average percent correct place of articulation identification across listeners for the 20 ms gating condition was 75.4% and 70.4% for the unprocessed-gated and minimum-phase conditions respectively. By comparison, the rate of correct place identification for 10 ms stimuli was 76.1% and 77.8% for the unprocessed-gated and minimum-phase conditions respectively. A decrease in performance for minimum-phase stimuli was observed only for the 20 ms long stimuli. In addition, correct place identification for the shorter stimuli was actually improved over the 20 ms gating duration. Because the 10 ms session was presented after the 20 ms gating condition for the five subjects that participated in both sessions, this may be due to some kind of training effect. Nevertheless, this differnece between gating durations appears to be small in magnitude for the unprocessed-gated stimuli.

These observations can be compared to results found by Tekieli and Cullinan (1979) who found 60% correct place of articulation identification in 10 ms long stimuli and 90% for 20 ms long stimuli. Similarly, Nearey (1992c) found 80% correct place identification performance in 13 ms stimuli and 70–90% for 18 ms stimuli.

Although it is possible to evaluate the significance of the effect of processing condition using a binomial test statistic, this is unsatisfactory for two reasons: the subjects in this experiment represent a random effect which must be taken into account (Clark, 1973)—*i.e.*, between speaker variation may create a large amount of overdispersion thereby reducing the accuracy of the significance levels (McCullagh and Nelder, 1989; Lindsey, 1999). In addition, because the unprocessed-gated and minimum-phase tokens originate from the same stimuli, the two sets of responses are not independent and it is therefore more appropriate to use some kind of paired-comparison test. For example, if one signal processing condition produces fewer errors than the other, but all of these errors are made in both conditions, a naïve binomial test statistic may conclude that the differences are not significant when indeed they are. This can be compared to a situation in which errors in both conditions are made on completely disjoint sets of stimuli which might be the case with two very similar, but *inconsistently* classified sets of stimuli (Ripley, 1996).

Instead, a McNemar test is performed (Fleiss, 1981). Responses to the both sets of stimuli A and B are relabelled as either right or wrong and the McNemar test determines whether a *correct* identification on an item in set A and an *incorrect* identification on the corresponding item in set B is more or less likely than the opposite.

The original McNemar statistic m has the form

$$m = \frac{(|n_A - n_B| - 1)^2}{n_A + n_B} \tag{2.2}$$

where n_A and n_B are the number of errors made in condition A not made in condition B and *vice versa*. Equation 2.2 is approximately distributed χ^2_1 under the null hypothesis H_0 when $n_A + n_B > 25$. In order to account for the random subject effect, the sum of these statistics across N subjects can therefore be compared to a χ^2_N distribution. However, a peculiarity of the McNemar test statistic is that it can only tell us if one condition is improved over the other, but not which one—*i.e.*, it is unsigned and therefore cannot represent the direction of improvement between condition A and condition B. This problem can be solved by taking the square root of the test statistic and setting its sign based on the direction of change between condition A and condition B. This new statistic is expected to have a standard normal distribution $\mathrm{N}(0, 1)$ under the null hypothesis.

An exact test of significance refers n_A to a binomial distribution $\mathrm{B}(n_A + n_B, \frac{1}{2})$ under the null hypothesis (Ripley, 1996). Table 2.1 on the following page gives these significance levels for each subject where the subscripts A and B refer to the unprocessed-gated and minimum-phase conditions respectively. In addition, the correct place-of-articulation identification rates are given for each subject in the experiment. From the table it can be seen that identification performance improves for unprocessed-gated stimuli for only one listener in the 10 ms gating condition. For 20 ms stimuli, all but one subject showed better performance for unmodified stimuli. Based on the significance levels under the column labeled α, significant decrements in performance were not observed for any of the subjects for the 10 ms stimuli. However, for the 20 ms stimuli three subjects showed a significant decrease in correct identification. The sum of the signed, square-root of the McNemar test statistics $\pm\sqrt{m}$ for each subject were applied to a one-tailed t-test with five degrees of freedom. This test found no significant decrement in performance for the minimum-phase stimuli in the 10 ms condition ($t = -1.54$; *n.s.*), but found a significant decrease for the 20 ms condition ($t = 2.57$; $\alpha < 0.05$). By comparison, these t-statistics were, in fact, very similar to those obtained using a simple (but presumably less powerful) one-tailed, paired-comparison t-test with 5 degrees of freedom: no significant decrease in correct identification for the 10 ms stimuli ($t = -1.54$; *n.s.*), but significant decreases in performance for the 20 ms stimuli ($t = 2.44$; $\alpha < 0.05$).

Thus temporal information in the form of Fourier phase appears to be important for the longer gating duration only. This supports the result obtained in Sec. 2.2.3 which found that the difference between static and dynamic representations of gated stop bursts was not significant at durations of 13.75 ms but were significant at 23.75 ms. It is also shown that a static power spectrum of the first 13.75 ms contains most of the perceptually relevant stop place information used by listeners. This cannot however be said for gated bursts of 23.75 ms in duration.

Table 2.2 on p. 45 gives the confusion matrices for both gating durations and for both the unprocessed-gated and minimum-phase signal processing conditions. The difference between the two signal processing conditions appears to be much larger for the 20 ms than for the 10 ms stimuli as was confirmed by the t-tests above. Much of this difference can be attributed to an increase in errors to alveolar stimuli (second row

	10 ms			20 ms		
subj.	pcc_A	pcc_B	α	pcc_A	pcc_B	α
A	...	...	...	0.56	0.53	0.26
B	0.70	0.74	0.91	0.71	0.73	0.81
C	0.67	0.72	0.91	...	...	...
D	0.83	0.81	0.19	0.89	0.81	0.00**
E	0.70	0.73	0.85	0.68	0.56	0.00**
F	0.82	0.84	0.81	0.86	0.81	0.02*
G	0.84	0.84	0.55	0.81	0.77	0.12

Table 2.1: Percent correct place of articulation identification for each subject. The columns labeled pcc_A and pcc_B refer to the correct identification rates for the unprocessed-gated and minimum-phase stimuli respectively. The significance levels under α are based on the comparison $n_A \sim \text{B}(n_A + n_B, \frac{1}{2})$ where $\text{B}(n, p)$ is the binomial distribution, and n_A and n_B are the number of errors made in the unprocessed-gated condition that are not made in the minimum-phase condition and *vice versa*.
$^*\alpha < 0.05$. $^{**}\alpha < 0.01$

in the bottom tables) which are often incorrectly categorized as bilabial. The stimulus example presented in Figs. 2.1 on p. 34 and 2.4 on p. 38 was one such stimulus and was perceived as alveolar 5/6 times for the unprocessed-gated condition and as bilabial 5/6 times for the minimum-phase condition. Although this pattern is somewhat evident in the 10 ms stimuli it is definitely not present to the same degree.

2.3.6 Perceptual modeling

As mentioned in Sec. 2.3.2, it may be that, although the minimum-phase decomposition removes all nonredundant Fourier phase information from the gated stimuli, the original stimuli may have already been nearly minimum-phase and, therefore, the magnitude of the temporal distortion was not sufficient to affect listeners' responses at the shorted gating duration. It may also be the case that changes in listeners' responses between the two signal-processing conditions are not directly related to the nature of the manipulations. In order to test these possibilities, listeners' responses were modeled using one of two sets of stimulus parameters as described in Sec. 2.2.1. MFCCs for both the unprocessed-gated and minimum-phase stimulus sets were measured in either one or two frames.

Listeners' responses to place of articulation were fit with a generalized linear model (GLM) in which the multinomial frequency data were treated as a Poisson loglinear process conditional on the total number of responses per stimulus (McCullagh and Nelder, 1989). This is typically done by minimizing the residual deviance statistic

$$D(y;\hat{\mu}) = 2\sum_i\sum_j\{y_{ij}\log(y_{ij}/\hat{\mu}_{ij}) - (y_{ij} - \hat{\mu}_{ij})\} \qquad (2.3)$$

using an iterative weighted least squares regression, where y_{ij} is the total number of category j responses to stimulus i, and $\hat{\mu}_{ij}$ is the estimate of this value based on the model parameters.

Alternatively, the regression coefficients and deviance statistic can be estimated by

	response: 10 ms duration					
stimuli	unprocessed			minimum-phase		
	lab	alv	vel	lab	alv	vel
labial	0.64	0.22	0.15	0.87	0.07	0.06
alveolar	0.02	0.94	0.04	0.15	0.75	0.11
velar	0.01	0.28	0.71	0.09	0.19	0.72
	response: 20 ms duration					
	unprocessed			minimum-phase		
	lab	alv	vel	lab	alv	vel
labial	0.71	0.19	0.09	0.78	0.11	0.10
alveolar	0.07	0.87	0.07	0.22	0.65	0.14
velar	0.03	0.28	0.68	0.12	0.20	0.68

Table 2.2: Confusion matrices for minimum-phase and unprocessed-gated stimuli pooled across subjects.

a single-layer neural network using maximum conditional likelihood fitting or SOFTMAX (Ripley, 1996). This is similar to the approach taken by Smits *et al.* (1996b) who also use a single-layer perceptron to model listeners' data. However, the model employed by Smits *et al.* assumes that the observed responses (or network outputs) are independent and a logistic analysis is performed on each category individually. The technique used here, however, is to fit surrogate Poisson loglinear models to the multinomial frequency data (Venables and Ripley, 1998) and condition the outputs (or predicted number of responses to each category for each stimulus) on the *total* number of responses.

In many respects, the approaches are, however, very similar. The analysis assumes that the proportion of responses to each place of articulation for each stimulus represents a normal *a posteriori* probability (NAPP) which further assumes that the three places of articulation can be represented by multivariate normal distributions in the space of continuous variates. If these multivariate distributions were known, *a priori* probabilities of class membership to each place of articulation could be determined based on the distance from each obsrvation from the category centers. In the absence of the exact category distributions, however, the GLM estimates normalized *a posteriori* probabilities on the basis of subject's classifications.

Like the model of Smits *et al.* (1996b), the NAPP model used in this analysis further assumes that all categories share a common covariance structure and that class boundaries are linear. Similar analyses have been performed using mulitnomial response data by Nearey and Assmann (1986); Andruski and Nearey (1992); Benkí (1998); and Hillenbrand and Nearey (1999).

Table 2.3 on the following page gives three goodness-of-fit statistics for the loglinear regression models. The residual deviance statistic D decreases for the dynamic parameters in both the 10 ms and 20 ms conditions. However, it is difficult to assess the significance of this improvement. Because the dynamic models contain twice as many parameters, it is guaranteed to be the case that they will have a lower residual deviance regardless of the actual significance of the additional parameters.

Two estimates that have been proposed for use in optimum model selection include

model	dev.	AIC	BIC
	10 ms gating condition		
static	2048.8	2104.8	2246.2
dynamic	1647.3	1759.3	2042.1
	20 ms gating condition		
static	2006.8	2062.8	2204.2
dynamic	1508.2	1620.2	1902.9

Table 2.3: Goodness of fit statistics for response models

the AIC (the Akaike information criterion) and BIC (Atkinson, 1981) which are both of the form

$$Q = D + \alpha q \phi \tag{2.4}$$

where q is the number of estimable parameters and ϕ is the dispersion parameter which is nominally equal to one in the case of multinomial data. For the AIC, $\alpha = 2$. For the BIC, $\alpha = \log(n)$, where n is the number of independent observations or $k(p-1)$ in the case of multinomial data, where k is the number of stimuli and p is the number of response categories. Therefore, the additional term $\alpha q \phi$ in Eq. 2.4 is intended to penalize larger models.

Both the AIC and the BIC in Table 2.3 decrease for the dynamic model in both gating durations. However, it is known that model selection on the basis of either of these statistics is necessarily biased towards the inclusion of unnecessary parameters (Davison and Hinkley, 1997). In addition, the AIC and BIC statistics will be underestimated in the context of overdispersion (*i.e.*, $\phi > 1$; Eq. 2.4) which is usually the case unless the model underlying the data can specified exactly, which is generally not possible (McCullagh and Nelder, 1989).

However, because we wish to address the issue of whether the dynamic spectral model is significantly closer to the true model underlying listeners' responses to these stimuli, it is necessary to use an unbiased model selection criterion which also gives an estimate of the *significance* of the improvement for the larger model. Several such procedures have been proposed (*e.g.*, Gumpertz and Pantula, 1989; Shao, 1993, 1996; Efron and Tibshirani, 1997). The approach taken in this analysis is a cross-validation procedure similar to the one used in Sec. 2.2.3.

Typically, in the cross-validation procedure, the dataset is divided into a *training* set S_t and an *assessment* set S_a. The model is fit to S_t, the parameters are then held fixed, and the fitted model is used to predict the responses in S_a. The goodness-of-fit statistic is then a function of the predicted and observed responses to the assessment set S_a which is completely independent from the observations used to actually train the model S_t.

In automatic classification experiments, training and test data would normally be selected from different sets of stimuli. However, there are two potential sources of random variation in this data: both *speakers* and *subjects*. If the data were split into training and assessment sets along speakers, there would still be a correlation between S_t and S_a because the same group of subjects produced responses to both sets of stimuli. Because of the presence of potentially two random effects, the data were partitioned in two stages: listeners were first divided into equal sets of three subjects each. In all,

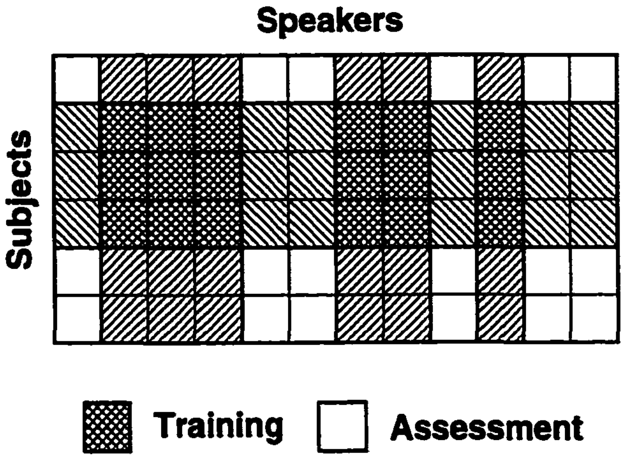


Figure 2.7: Illustration of cross-validation procedure. Three subjects are randomly selected and assigned to the listener training set (filled rows). Six speakers are then assigned to the talker training set (filled columns). The model was then fit to the data represented by the intersection of filled rows and columns and was then used to predict the data represented by the unfilled squares.

$\binom{3}{6} = 20$ such partitionings are possible and therefore an exhaustive sampling of all possible combinations was computationally feasible. However, the datasets were then partitioned again into two orthogonal sets of six speakers each. Essentially, the data were split into four nonoverlapping sets: S_{tt}, S_{ta}, S_{at}, and S_{aa}, where the first subscript refers to the subject partitions, the second refers to the speaker partitions, and each label refers to the intersection of the corresponding subsets. The model can then be fit to S_{tt} and used to predict responses to S_{aa}. Goodness-of-fit statistics on the basis of these predictions, when compared to observed responses in S_{aa}, are assured of being unbiased, since there can be no correlation between the two sets.

Figure 2.7 illustrates the cross-validation procedure used here. At each subject partioning, the responses from three listeners (filled rows) to stimuli from six randomly selected speakers (filled columns) are used to estimate the parameters of a GLM—*i.e.*, the model is fit to the data represented by the intersection of filled rows and columns. Once this model is obtained, its parameters are held fixed and it is used to predict responses to the stimuli that ***were not*** used in the fitting stage represented by the unfilled squares.

For each of the 20 listener partitionings, 25 more speaker partitionings were randomly generated for a total of 500 such iterations. Note that at each iteration, only one quarter of data is used in each of the assessment S_{aa} and training S_{tt} sets and thus half of the data is discarded (S_{at} and S_{ta}). Nevertheless, we are assured that the data used to estimate the model are completely independent from the data with which the model is evaluated. It has additinoally been shown that using a relatively small set of observations in the model estimation stage results in consistent model selection in cross-validation (Shao, 1993).

If it is indeed the case that the larger model is significantly closer to the true model, then we expect that 95th percentile of the paired difference between the goodness-of-fit statistics to indicate improvement—*i.e.*, the significance level of the difference in goodness-of-fit scores Δ is

$$\alpha = \frac{\#\{\Delta_r^A - \Delta_r^B < 0\}}{R} = \frac{1}{R}\sum_{r=1}^{R} I\{\Delta_r^A - \Delta_r^B < 0\} \qquad (2.5)$$

where A and B refer to the dynamic and static models respectively, R is the total number of trials in the cross-validation test, and $I\{E\}$ is the indicator function of event E: $I\{E\} = 1$ iff E is true, 0 otherwise. Therefore, a significant difference would be indicated by an improvement in the goodness-of-fit statistic in more than 95% or 475 trials (Efron and Tibshirani, 1993; Davison and Hinkley, 1997).

For the goodness-of-fit statistic itself, Shao (1996) recommends using Pearson's X^2

$$X^2 = \sum_i \sum_j \frac{(y_{ij} - \hat{\mu}_{ij})^2}{\mathrm{var}(\hat{\mu}_{ij})} \qquad (2.6)$$

where $\mathrm{var}(\hat{\mu})$ is the variance of the estimate $\hat{\mu}$ which is nominally

$$\mathrm{var}(\hat{\mu}_{ij}) = \hat{\mu}_{ij} \qquad (2.7)$$

for loglinear data. However, the Pearsons X^2 statistic is easily inflated by very small values of $\hat{\mu}$ which often occur in multinomial response data which are generally very sparse—*i.e.*, mostly zeros. It also underestimates the effects of overdispersion in which the empirical variance is larger than that estimated by Eq. 2.7.

Instead two other measures of goodness of fit are used. The first is the percentage of modal agreement (pma) between the predicted and observed responses (Hillenbrand and Nearey, 1999). This measures the proportion of observed responses that show the same modal category as that predicted by the model. This can be defined as

$$\mathrm{pma} = \frac{1}{N}\sum_{i=1}^{N} I\left\{ \arg\max_j(\hat{\mu}_{ij}) = \arg\max_j(y_{ij}) \right\} \qquad (2.8)$$

where N is the total number of stimuli in the assessment set S_{aa}. Smits *et al.* (1996b) also report pma goodness-of-fit statistics for their models.

The second measure used is the root-mean-squared error (rms) between the observed and predicted response probabilities:

$$\mathrm{rms} = \left[\frac{1}{NK} \sum_{i=1}^{N} \sum_{j=1}^{K} (y_{ij} - \hat{\mu}_{ij})^2 \right]^{\frac{1}{2}} \qquad (2.9)$$

where K is the number of response categories. The square of this value is proportional to the residual sum of squares in a *linear* regression and therefore gives equal weight to all the prediction errors regardless of their estimated variance (Eq. 2.7). Although the model was fit using a *generalized* linear regression, it has been shown that a simple linear regression will produce similar results if the distribution of the observed stimulus parameters is approximately multivariate normal $N_p(\mu_i, \Sigma)$ with common covariance matrices Σ for each category (Ripley, 1996). Hillenbrand and Nearey (1999) use a similar statistic based on Pearson's coefficient of correlation ρ between the observed and predicted response probabilities.

For the 10 ms gating condition, the percent modal agreement was better in the dynamic model for only 71.4% of the trials (*cf.*, Eq. 2.5 on the preceding page). Unlike the AIC and BIC tests given in Table 2.3, this suggests that the dynamic spectral model is not significantly closer to the true model underlying listeners' responses to 10 ms gated stop bursts. For the 20 ms gating condition, however, the pma was higher in 95.8% of the iterations, suggesting that dynamic spectral information is indeed exploited by listeners at longer gating durations.

The dynamic model had a lower rms than the static model in only 68.8% of the trials for the 10 ms condition and 91.2% in the 20 ms condition indicating only marginal significance for differences in the longer gating duration.

In summary, the unbiased crossvalidation procedure for evaluating the difference between minimum-phase and original-gated stimuli showed significant differences for the 20 ms stimuli only indicating that there is no phonetically relevant phonetic information present in gated bursts of 10 ms duration.

2.4 Relationship to VOT

An explanation for the results in the perceptual experiment in Sec. 2.3 can perhaps be found by examining specific temporal properties in these stimuli. For example, it is clear that the stimulus presented in Fig. 2.4 on p. 38 contains at least one full glottal pulse. The group delay response of the allpass component in Fig. 2.6 on p. 40 indicates that, in addition to high-frequency energy above 4 kHz, a large amount of low-frequency energy below 800 Hz is shifted backward in time towards the onset of the stimulus. This low-frequency energy can be largely attributed to the short period of voicing that is present in the unprocessed stimulus. It is possible that the change in temporal relationship between the vocalic and burst portions has contributed to the increase in bilabial responses to the minimum-phase stimulus.

Figure 2.8 on the next page shows histograms of voice onset times for the six consonants used in this experiment. The figure shows that most of the VOTs for the voiced stops are less than 20 ms, while most of the voiceless VOTs are above this value. Therefore, it may be the case that many of the errors in correct identification observed in this experiment are due to the temporal smearing of the burst and vocalic portions in the minimum-phase condition. If this hypothesis is correct, then we should expect that there would be no difference in correct identification performance between the two signal processing conditions for the *voiceless* stimuli as the vast majority of these VOTs are much longer than the stimulus durations themselves and therefore no voicing should be found in the 20 ms gated stimuli.

Table 2.4 on the following page gives separate confusion matrices for the voiced and voiceless stimuli. The table shows that the pattern of errors to alveolar stimuli observed in Table 2.2 on p. 45 is not evident for the voiceless stops even at the 20 ms gating duration—*i.e.*, the number of bilabial responses to [d] in the minimum-phase conditions is much larger than that for [t]. This would seem to indicate that many of the errors to minimum-phase stimuli in the 20 ms condition may be related to VOT.

Table 2.5 on p. 51 gives the significance levels for the decrease in correct identification performance for the minimum-phase condition for both voiced and voiceless stimuli for each subject. There appears to be no significant effect for processing condition for any of the subjects for the voiceless stimuli in either the 10 ms or 20 ms gating conditions. For the voiced stimuli however, four of the six subjects showed significant

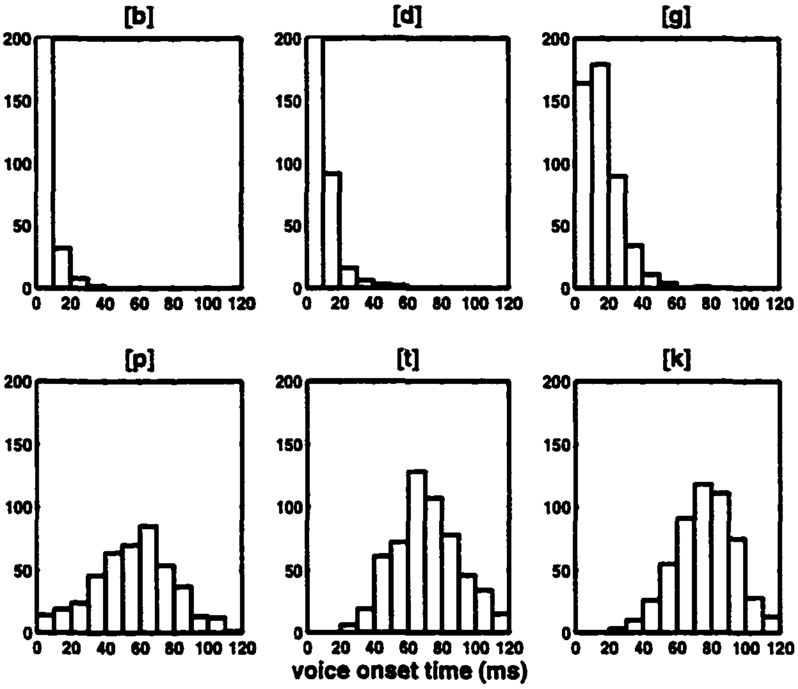


Figure 2.8: Histograms of voice onset times for the six consonants.

stimuli	response: 10 ms duration					
	unprocessed			minimum-phase		
	lab	alv	vel	lab	alv	vel
p	0.52	0.30	0.19	0.80	0.13	0.08
t	0.00	0.98	0.02	0.08	0.81	0.11
k	0.01	0.26	0.73	0.07	0.17	0.75
b	0.75	0.14	0.10	0.94	0.02	0.04
d	0.05	0.90	0.05	0.21	0.68	0.10
g	0.01	0.30	0.69	0.11	0.20	0.68

	response: 20 ms duration					
	unprocessed			minimum-phase		
	lab	alv	vel	lab	alv	vel
p	0.54	0.30	0.16	0.67	0.19	0.14
t	0.02	0.90	0.08	0.05	0.78	0.17
k	0.03	0.26	0.71	0.07	0.18	0.74
b	0.88	0.09	0.03	0.90	0.04	0.06
d	0.11	0.83	0.06	0.38	0.52	0.10
g	0.03	0.31	0.66	0.16	0.23	0.61

Table 2.4: Separate confusion matrices for voiced and voiceless stimuli.

	10 ms			20 ms		
subj.	pcc_A	pcc_B	α	pcc_A	pcc_B	α
			voiceless stops			
A	···	···	···	0.52	0.48	0.25
B	0.67	0.74	0.96	0.67	0.78	1.00
C	0.67	0.68	0.67	···	···	···
D	0.84	0.83	0.43	0.85	0.86	0.65
E	0.67	0.78	0.99	0.63	0.62	0.44
F	0.78	0.84	0.98	0.85	0.84	0.41
G	0.82	0.85	0.84	0.78	0.80	0.75
			voiced stops			
A	···	···	···	0.60	0.59	0.45
B	0.73	0.74	0.63	0.75	0.69	0.11
C	0.68	0.75	0.94	···	···	···
D	0.83	0.78	0.18	0.93	0.76	0.00^{**}
E	0.73	0.69	0.22	0.74	0.51	0.00^{**}
F	0.85	0.83	0.35	0.88	0.78	0.01^{*}
G	0.86	0.83	0.20	0.84	0.74	0.01^{*}

Table 2.5: Percent correct identification for voiced and voiceless stops.
$^{*}\alpha < 0.05$. $^{**}\alpha < 0.01$

effects for processing condition in the 20 ms gating condition. The modified McNemar statistics $\pm\sqrt{m}$ for the voiced and voiceless stimuli were subjected to one-tailed t-tests and it was found that the signal processing condition had no significant effect on correct identification performance for any of the voiceless stimuli ($t = -2.69$; *n.s.* and $t = -0.63$; *n.s.* for the 10 ms and 20 ms conditions respectively). However, a significant effect was found for the voiced stimuli in the 20 ms condition ($t = 0.78$; *n.s.* and $t = 3.50$; $\alpha < 0.01$ for the 10 ms and 20 ms conditions respectively). One-tailed, paired-comparison t-tests produced similar results: no significant change for voiceless stops ($t = -2.58$; *n.s.* and $t = -0.53$; *n.s.* for 10 ms and 20 ms durations respectively) and significant result only for the 20 ms duration in the voiced stops ($t = 0.59$; *n.s.* and $t = 3.51$; $\alpha < 0.01$ for the 10 ms and 20 ms durations respectively).

Therefore, subjects had significantly more difficulty in identifying the place of articulation of 20 ms long temporally distorted, gated voiced stops. Otherwise, there were no significant changes in performance between the two signal processing conditions. Based on the distribution of VOTs in Fig. 2.8 on the preceding page, this would seem to indicate that distortion of the temporal relationship between the burst and vocalic segments has a deleterious effect on perception.

This hypothesis was further verified by modeling listeners' responses in a multinomial loglinear regression using the statistical procedure described above. However only the voiceless consonants in the 20 ms gating condition are considered. In contrast to the results obtained for both voiced and voiceless bursts pooled together, the dynamic representation had a larger percent modal agreement in only 69.6% of the trials while the static rerpresentation had a larger rms in 64.2% of the trials. This lends further evidence for the importance of VOT perception in gated bursts.

2.5 Conclusions and discussion

The first experiment in Sec. 2.2 showed that some spectrally dynamic, discriminant information is present in short gated stop bursts of 23.75 ms duration. Although automatic classification via LDA at shorter durations was better using dynamic spectral properties, it was shown that this difference was not significant. Therefore, on the basis of these results, we had expected that human listeners might be able to exploit dynamic spectral information in 20 ms long gated stop bursts but that isolated segments shorter than this might not contain phonetically relevant dynamic information.

Section 2.3 confirmed this hypothesis: listeners did not exhibit significantly different categorization behavior in the 10 ms gating duration condition. However, 20 ms minimum-phase stimuli were identified with significantly more errors than in the unprocessed-gated condition. By modeling listeners' responses using either static or dynamic spectral features, it was confirmed that listeners did not rely on dynamic spectral information at the shorter duration.

However, it was subsequently suggested that the difference in error rates between the two gating durations may be related to the onset of voicing and whether short VOT glottal epochs were present in the gated stimuli. It was found that there was no significant difference in listeners' ability to classify temporally distorted voiceless bursts at the 20 ms gating condition, but that voiced bursts showed significantly more errors in the minimum-phase condition. A comparison of the cross-validated goodness-of-fit statistics found that there was no significant improvement associated with dynamic spectral features for voiceless stop bursts.

Based on measurement of VOT from voiced and voiceless stops, it was concluded that a static spectral representation is adequate for isolated stop bursts—at least up to a maximum duration of 20 ms—and that an important temporal cue is the relative onset of voicing. No significant dynamic spectral information was found to be exploited by listeners in stimuli that contained only the release burst without any of the following vocalic segment. It is well known that VOT is an important secondary cue to place of articulation (Lisker and Abramson, 1964; Oden and Massaro, 1978; Kewley-Port, 1982; Benkí, 1998). These results show however, that the burst and subsequent vocalic segments must be held distinct in perception. This is in contradiction with previous analyses in which the distinction between burst and vocalic segments was blurred (*e.g.*, Blumstein and Stevens, 1980).

An ideal test of this hypothesis would be to present listeners with the minimum-phase component of isolated bursts. However, this introduces the additional temporal cue of burst duration which has also been shown to have a significant effect in subject responses in the perception of segmented bursts (Bonneau, Djezzar, and Laprie, 1996).

These results refute the claim made by Blumstein and Stevens (1980) that static spectra of the first 10–20 ms following consonant onset is an invariant cue to place of articulation perception *irrespective* of VOT, or even whether the burst itself is present or not. Conversely, the absence of an effect for waveform envelope information in stimuli with longer VOTs refutes the claim that such purely temporal information can be used to discriminate between different places of articulation (Cole and Scott, 1974b). However, it lends credibility to the view that the burst, isolated from the subsequent vocalic portion can be treated as a stationary acoustic signal in that temporal modulations can effectively be ignored for the purposes of phonetic perception. Spectral measurements of isolated release bursts have typically ignored dynamic spectral properties (*e.g.*, Fischer-Jørgensen, 1954; Winitz, Scheib, and Reeds, 1972; Dorman, Studdert-Kennedy, and Raphael, 1977; Jongman and Miller, 1991). Based on the present results, this view seems defensible. However, studies that assign a fixed window duration in burst analyses must be treated with caution lest they include portions of the subsequent voiced segment (*e.g.*, Halle, Hughes, and Radley, 1957; Cole and Scott, 1974b; Zue, 1976).

It is interesting to review the literature on spectrally dynamic representations of stop consonant onsets and evaluate their ability to accommodate the results found in the present experiment. Kewley-Port (1983) and Kewley-Port and Luce (1984) proposed that stop place of articulation could be characterized by dynamic spectral templates. These templates were able to incorporate F_1-cutback which is correlated with VOT and has therefore been found to be partially distinctive for velar stops. However, they also suggested that discrimination of [b] and [d] could be performed on the basis of the first 20 ms alone. It was shown in Sec. 2.3.5 that most of the confusions in minimum-phase stimuli were actually between [b] and [d] and therefore these stops would also have to be distinguished on the basis of VOT (*cf.*, Searle, Jacobson, and Kimberley, 1980).

The discrimination between bilabials and alveolars was addressed by Lahiri *et al.* (1984), who suggested that the *relative* change in global spectral energy between the burst and the onset of the vocalic segment could be used to classify stop consonants. They suggested that alveolars were characterized by a greater change in low-frequency energy between the burst and vocalic portions than for bilabials—*i.e.*, there was a relative increase in low-frequency energy at the onset of voicing for alveolar stops. This theory corresponds most closely to the results found in these experiments: alveolars with VOTs shorter than the duration of the gating condition were perceived as bilabials when low-frequency energy was shifted towards the stimulus onset.

Chapter 3

The perception of spectrally reduced prevocalic stop consonants

3.1 Introduction

The purpose of the present chapter is to evaluate the relative perceptual importance of gross *vs.* detailed spectral cues in the perception of prevocalic stop consonants. Similar to Chap. 2, this comparison is made by studying the effects of specific types of distortion on the perception of stop-consonant place of articulation. In contrast to the previous chapter, however, the distortion is primarily restricted to the *spectral* properties of naturally produced speech rather than *temporal* aspects.

Smits, ten Bosch, and Collier (1996a) have defined detailed spectral cues as those which must be represented using a frequency resolution no broader than 500 Hz. In the perception of prevocalic stop consonants, such cues may include the burst peak frequency (*e.g.* Liberman, Delattre, and Cooper, 1952; Zue, 1976) as well as formant transitions for F_2 (*e.g.*, Cooper *et al.*, 1952; Liberman *et al.*, 1954; Delattre *et al.*, 1955; Kewley-Port, 1982) and F_3 (*e.g.*, Harris *et al.*, 1958; Hoffman, 1958). Locus equations also fall into this category (Nearey and Shammass, 1987; Sussman *et al.*, 1998).

While these acoustic cues are defined purely in terms of narrowly localized spectral features, *gross* spectral cues can be broadly represented by overall spectral shape requiring much less spectral detail. For example, it has been proposed that the gross spectral tilt and relative spectral compactness of the first 10–20 ms of the stimulus following the onset of the release burst is a sufficient cue to stop consonant place of articulation (Blumstein and Stevens, 1980). Likewise, it has also been proposed that the *relative change* in spectral tilt from the burst to the onset of voicing is a sufficient cue for discrimination between bilabials and alveolars (Lahiri, Gewirth, and Blumstein, 1984). The gross spectro-temporal envelope over the first few tens of milliseconds has also been suggested (Kewley-Port, 1983). A common characteristic of gross spectral cue theories is that the spectro-temporal envelope is preserved as a holistic, indivisible unit, and as a consequence, locally defined properties can interact in relatively arbitrary ways. For example, although spectral slope is determined primarily by the relative amplitudes of prominent spectral peaks, the absolute frequencies of individual peaks are not considered to be independent perceptual cues (*e.g.*, Lahiri *et al.*, 1984). In contrast, detailed cue theories treat specific acoustic features, such as burst peak or formant frequency, as relatively distinct perceptual properties.

Three different experimental paradigms have been employed to evaluate the relative importance of gross *vs.* detailed cues as acoustic correlates to stop place of articulation: (a) conflicting-cue perception experiments, (b) statistical modeling of listeners' responses, and (c) automatic classification. In the first type of experiment, subjects are asked to classify stimuli in which gross and detailed features cue conflicting places of articulation to determine which set of acoustic cues dominates listeners' responses. Using synthetically produced stops, Blumstein, Isaacs, and Mertus (1982) found that the majority of listeners' identifications were appropriate for the place of articulation specified by the formant frequencies. Walley and Carrell (1983) also manipulated formant transitions and onset spectral shape parameters independently, and found that subjects classified conflicting cue stimuli primarily on the basis of formant transition information. Similarly Lindholm *et al.* (1988) found that normal-hearing listeners responded appropriately to formant transitions more often than to the overall spectral tilt. Although hearing-impaired listeners showed the opposite effect, their responses were much less consistent. Similarly, Dorman and Loizou (1996) also found that listeners' responses were determined primarily by the place of articulation as specified by formant transitions. The stimuli used in this experiment were derived from naturally produced tokens and were manipulated such that the relative change in spectral tilt between the burst and the onset of voicing specified a place of articulation that conflicted with detailed spectral features.

Another approach has been to model listeners' responses using explicit detailed or gross-cue hypotheses. For example, Krull (1990) compared place confusions in subjects' identifications to Euclidean distances between place category centers in either gross or detailed feature spaces. She found that relative distances based on onset formant frequencies of F_2–F_4, when combined with burst duration, were better able to predict listeners' confusions than gross spectral shape which were represented by filterband spectra. Similarly, Smits, ten Bosch, and Collier (1996b) found that detailed cues, such as formant transitions, gave a better fit to listeners' responses to cross-spliced, naturally produced stimuli in which the burst and vocalic segments were concatenated from conflicting places of articulation. However, in contrast to these results, Nossair and Zahorian (1991) found that global spectral shape properties in the form cepstral coefficients provided better stop classification performance than formant frequencies in an automatic classification experiment.

The approach taken in this chapter will be to examine the perception of *spectrally distorted* speech under the assumption that the reduced frequency resolution will eliminate detailed frequency properties while preserving gross spectral shape. If it is indeed the case that detailed spectral cues are perceptually more important in the identification of stop place of articulation, then we expect correct identification performance to degrade quickly in the context of reduced frequency resolution.

In their work on the simulation of speech perception by recipients of cochlear implants, Shannon *et al.* (1995) have shown that normal-hearing listeners can be trained to identify stop consonant place of articulation from only four channels of amplitude modulated (AM) noise (hereafter referred to as "noise-vocoded speech"). The stimuli used in these experiments were produced by modulating the amplitude of bandlimited noise by the outputs of a bank of bandpass analysis filters. Although the waveform envelope of each subband is retained, frequency information within each channel is completely distorted. The spectral resolution of the resulting stimuli is therefore proportionately reduced by the bandwidth of the channel filters.

An example of stimuli produced in this manner is given in Fig. 3.1 on the following page. Figure 3.1(a) shows a wideband spectrogram of the syllable [dok] as produced by a female speaker, while the remaining spectrograms show the effects of noise vocoding at three different bandwidths: 500 Hz (b), 1000 Hz (c), and 2000 Hz (d). Figure 3.1(a) shows falling F_2 and F_3 formant transitions characteristic of an alveolar place of articulation (Cooper *et al.*, 1952; Liberman *et al.*, 1954; Harris *et al.*, 1958). It can be also be seen that the second formant originates in the vicinity of 1800 Hz which has been proposed as a characteristic F_2 locus for alveolars (Delattre *et al.*, 1955; Sussman *et al.*, 1998). Although these facts are evident to a certain degree at 500 Hz bandwidth [Fig. 3.1(b)], very little is discernible regarding the duration or extent of the relevant formant transitions in the remaining spectrograms.

Correct phoneme identification in the absence of such detailed spectral information prompted Shannon *et al.* (1995) to conclude that speech perception can be achieved largely on the basis of temporal information alone—*i.e.*, the only spectral information available to listeners is the relative energy in each subband. It should be noted, however, that although high correct recognition rates were obtained for place of articulation, subjects in this experiment had a total of 8–10 hours experience with such stimuli. This can be compared with results reported by Lindholm *et al.* (1988) who found that place perception by hearing-impaired listeners corresponded more closely to overall spectral tilt. Although it is *possible* to identify stop consonants on the basis of gross spectral shape given enough experience, it does not necessarily imply that listeners do so under normal hearing conditions.

Unlike the examples given in Fig. 3.1 on the next page, however, Shannon *et al.* (1995) used unequal channel bandwidths with filter cutoffs at 800 Hz, 1500 Hz, and 2500 Hz. As these stimuli were lowpass filtered at 4 kHz, filter bandwidths were therefore 800 Hz, 700 Hz, 1000 Hz, and 1500 Hz. If we focus attention on spectral distortion in the region of F_2, which has been postulated as the most important of the detailed perceptual cues (Liberman *et al.*, 1967), Shannon *et al.*'s four-channel noise-vocoded stimuli are most similar to Fig. 3.1(c).

Similar results have been reported elsewhere. Dorman, Loizou, and Rainey (1997) found that the perception of place of articulation did not improve significantly when the number of channels was increased beyond six with bandwidths of 187 Hz, 304 Hz, 493 Hz, 801 Hz, 1301 Hz, and 2113 Hz. In the current example, F_2 would straddle the 801 Hz and 1301 Hz bandwidth filters [*cf.*, Fig. 3.1(c) on the following page]. However, there was a significant decrease in correct place identification at wider bandwidths. Dorman *et al.* also demonstrated similar effects using amplitude-modulated sinusoids instead of bandlimited noise which was intended to emulate the processing strategy used by many cochlear implants (Dorman and Loizou, 1998). This can be compared with a very common situation in naturally produced, voiced speech with a very high fundamental frequency f_0, in which spectra are discretely sampled at harmonics of the periodic source—*i.e.*, voiced speech can be viewed as the sum of equally spaced AM sinusoids. It has been shown that an f_0 as low as 216 Hz is sufficient to eliminate the spectral peaks associated with the formants of some vowels without a noticeable decrease in intelligibility [see de Cheveigné and Kawahara (1999) for an overview of this effect].

It would appear then that listeners in experiments by Shannon *et al.* (1995) and

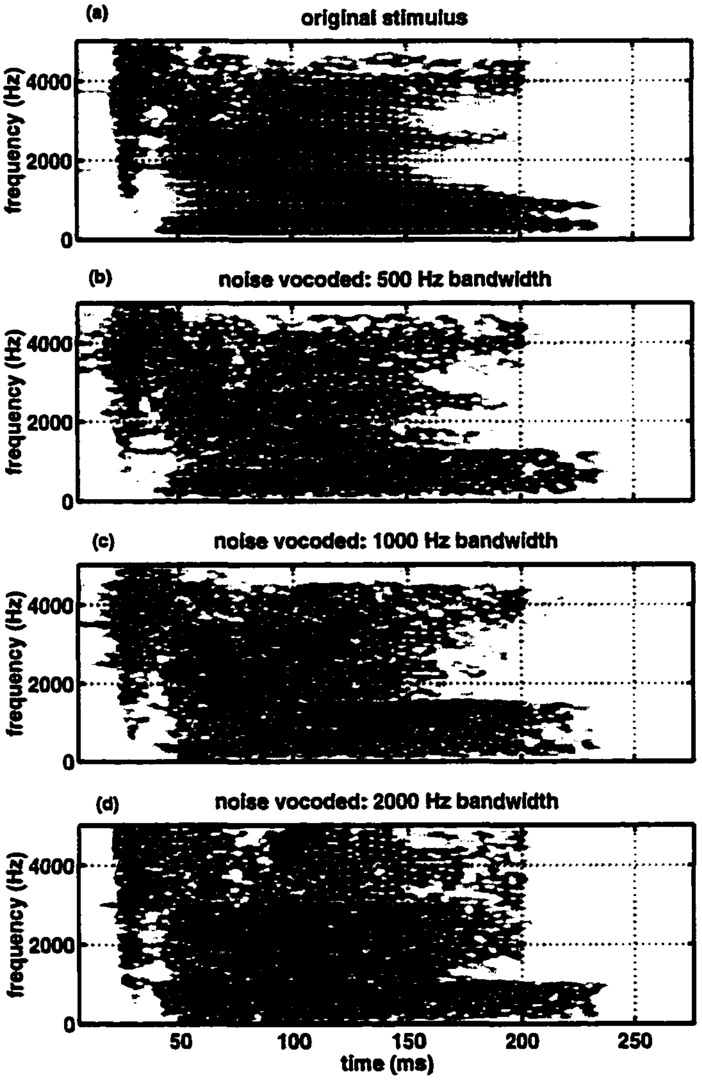


Figure 3.1: Sample spectrograms of noise-vocoded speech. Figure (a) shows the original token [dok] as produced by a female speaker. Figures (b), (c), and (d) show spectrograms of the same utterance after noise vocoding at three different channel bandwidths. The spectrograms were prepared using a 6.3 ms Hamming window with an approximate equivalent rectangular bandwidth of 180 Hz.

Dorman *et al.* (1997) are not able to rely on *detailed* spectral information such as F_2 or F_3 formant transitions. It could also be argued that these signal manipulations preserve *gross-spectral* cues such as overall spectral tilt, as well as the relative change in this property over time. Broadly defined spectral cues are potentially well preserved in noise-vocoded speech. For example, in the case of Shannon *et al.*'s four-channel cochlear implant simulation, it is likely that the energy in the lowest and highest frequency channels as well as the relative difference between the two mid-frequency channels is sufficient to represent gross spectral tilt. Likewise, the presence of relatively high energy in just one of the mid-frequency channels may preserve the property of spectral "compactness". Evidence that listeners are able to perceive stop place of articulation from such spectrally distorted stimuli lends support to such gross cue theories of speech perception.

However, very different results regarding the robustness of human speech perception in the context of reduced frequency resolution have been obtained using other signal processing techniques. In a perceptual experiment in which the signal itself was amplitude modulated by bandlimited noise, Boothroyd *et al.* (1996) found that a reduction in spectral resolution broader than 250 Hz produced a significant adverse effect on perception. Nevertheless, it was found that consonant perception was more robust than vowel perception. Two other similar experiments have shown that a reduction in spectral resolution beyond a critical bandwidth (or approximately $\frac{1}{3}$ octave above 400 Hz) is sufficient to produce a decrease in correct phoneme identification [at 1600 Hz, one critical bandwidth has been measured at 240 Hz (Zwicker, Flottorp, and Stevens, 1957)]. In one experiment, Celmer and Bienvenue (1987) manipulated the spectra of overlapping analysis frames directly, while in the second, ter Keurs, Festen, and Plomp (1992) lowpass filtered the log-frequency power spectra of overlapping segments. These two experiments must be treated with some caution however, as both implemented an overlap-add signal-processing technique which is not strictly appropriate for either dynamic or nonlinear filtering (Allen and Rabiner, 1977). Nevertheless, they show that listeners' ability to understand speech may depend on the presence and detectability of more detailed spectral cues after all.

In contrast to the noise-vocoding experiments described by Shannon *et al.* (1995) and Dorman *et al.* (1997), subjects in experiments using alternative signal-processing techniques were not extensively familiarized with the distorted stimuli before response data was collected. It may be the case that long exposure to spectrally distorted stimuli induces listeners to adopt an alternative set of acoustic features as perceptual correlates. For example, in Dorman *et al.*'s study, subjects listened to the 9-channel simulations first, thereby allowing them to adapt their perception to increasingly distorted stimuli. In simulations of electric hearing, this is not unwarranted; new recipients of cochlear implants are also given extensive training. However, it does not directly address the issue of how normal-hearing listeners perceive speech.

While it can be assumed that gross spectral properties such as global spectral shape is roughly preserved in noise-vocoded or otherwise spectrally degraded speech, it is not clear how much detailed spectral information is preserved by these signal manipulations. As mentioned above, Smits *et al.* (1996a) have defined detailed spectral properties as precisely those that must be represented using a frequency resolution no broader than approximately 500 Hz. This claim can be justified in a number of ways; for male voices, we typically expect roughly one formant for every 1 kHz in the dynamic spectral range of the speech signal, and fewer for female or children's voices. A common method for measuring formant frequencies in voiced speech is solving for the roots of linear predictor coefficients (LPC) in which each formant is represented by a single pole. Because two coefficients are needed for each pole, the LPC analysis requires roughly a 1000 Hz/2 = 500 Hz resolution for the coefficients to be completely recoverable from their spectral representation[1]. This is only a rough estimate of the spectral resolution that might be needed to represent formants for the average male speaker. However, because LPC analysis encodes not only formant frequencies, but also their bandwidths, it is not known how much frequency information alone might be recoverable from spectra that are spectrally degraded beyond 500 Hz.

In conclusion, there appears to be two conflicting sets of results in the literature on reduced spectral perception: on the one hand Dorman *et al.* (1997) claim that perception does not improve significantly when spectral resolution in the vicinity of F_2 is narrower than approximately 1000 Hz. This is also supported by Shannon *et al.*'s (1995) results. On the other hand, Boothroyd *et al.* (1996); Celmer and Bienvenue (1987); and ter Keurs *et al.* (1992) suggest that perception is significantly degraded beyond a spectral resolution in the order of 250 Hz. Differences in these results may reflect the effect of listener adaptation to spectrally distorted speech. In addition, while it is assumed that gross spectral properties do not require a fine spectral resolution, it is not known how much detailed frequency information is recoverable from spectrally degraded speech. These two questions are therefore addressed in the present chapter:

1. What minimum spectral resolution is required to fully represent detailed spectral cues in naturally produced speech? and

2. What minimum spectral resolution is required for the perception of stop consonant place of articulation for normal-hearing listeners without prior adaptation?

Section 3.2 attempts to answer the first question. Section 3.3 evaluates listeners' ability to identify stop consonants from noise-vocoded stimuli. Subjects' responses are then modeled using explicit theories of speech perception. Finally, Sec. 3.4 discusses any conclusions that can be drawn from this work regarding the nature of detailed *vs.* gross acoustic cues in stop consonant place perception.

3.2 Recoverability of gross and detailed spectral cues

Before a comparison can be made between gross *vs.* detailed acoustic cues in the perception of noise-vocoded speech, the issue of how detailed spectral features might be represented in such spectrally degraded stimuli must be addressed. Unfortunately, the measurement of formant frequencies is already relatively difficult in clean, natural speech and many of the *ad hoc* strategies that are employed to isolate them must be factored out of the larger issue of how they might be perceived in the context of reduced frequency resolution.

To this end, the *potential* recoverability of gross and detailed spectral features from noise-vocoded speech is evaluated by comparing the behavior of detailed spectral measurements to a completely random process using an appropriate nonparametric test of

[1] A more technical explanation could be that an Nth order LPC analysis estimated using the Levinson-Durbin recursion algorithm requires at least N autoregressive coefficients for a unique solution. Because the autoregressive series is the inverse Fourier transform of the power spectrum of the original signal, at least an N-point discrete Fourier transform is needed to generate the necessary power spectrum—*i.e.*, a minimum resolution of f_s/N, where f_s is the sampling rate. Therefore, with a sampling rate of 10 kHz, a minimum spectral resolution of 500 Hz is required for a 10th order LPC analysis representing five poles.

significance. The stimuli used in this experiment are described in Sec. 3.2.1. The signal processing is then described in Sec. 3.2.2. In Sec. 3.2.3, the procedure whereby spectral features were measured is given and the nonparametric test is explained in Sec. 3.2.4. The results are presented in Sec. 3.2.5 and, finally, the *robustness* of estimated detailed and gross spectral features is evaluated in an automatic classification task in Sec. 3.2.6.

3.2.1 Stimuli

The stimuli used in this experiment are similar to those described in Sec. 2.3.1. CVC syllables were recorded from 12 native speakers of Western Canadian English (five males and seven females). The initial consonant was one of the six stops [p], [t], [k], [b], [d], and [g], while the final consonant was always [k]. The medial vowel was one of [i], [ɪ], [e], [ɛ], [æ], [ʌ], [ɒ], [o], [ʊ], and [u]. However, because of the additional difficulties associated with the measurement of formants from the aspiration segment of voiceless stops, only the voiced stops [b], [d], and [g] are considered in this experiment. In total 3 × 9 × 12 = 324 syllables were used. All syllables were processed by the noise vocoder described in the next section.

3.2.2 Noise vocoding

Unlike the stimuli used in similar experiments involving noise vocoding which used analogue IIR filters to decompose the speech signal (*e.g.*, Shannon *et al.*, 1995; Dorman *et al.*, 1997), the stimuli in the present experiment were processed by a polyphase/FFT implementation of a high-order, linear-phase, uniform-bandwidth, complex FIR filterbank. This had several advantages. Because the filters were of high-order (134 × M where M is the number of channels in the filterbank, including both positive and negative frequencies), the subband decomposition ensured minimal spectral aliasing when the individual channels were critically downsampled. For example, with 16 channels at 16 kHz sampling rate—*i.e.*, 1 kHz bandwidth—the filter skirts had a steepness of approximately 2700 dB/octave. This is expected to be sufficient to eliminate any audible acoustic information outside the nominal passband (Warren and Bashford, 1999). It also ensured that the stimuli could be reconstructed from the subbands with minimal spectral distortion. In addition, because the filters were linear phase, phase distortion within each channel was minimized.

Instead of processing the stimuli with such a bank of FIR filters directly, the signals were decomposed via a polyphase/FFT realization of a uniform filterbank (Akansu and Haddad, 1992). The implementation is illustrated in Fig 3.2 on the next page and the MATLAB program used to generate the stimuli is presented in the appendix at the end of the chapter. Essentially, the polyphase decomposition of the original signal was filtered by the polyphase decomposition of a single lowpass FIR filter. Each frame across the M channels was then Fourier transformed to produce a bank of M frequency-shifted, critically sampled, complex subbands. Each subband was then substituted with signal correlated noise by randomly modulating the sign of the real component. Although the polyphase/FFT analysis produced complex waveforms, only the positive frequency subbands were modulated in this manner—*i.e.*, negative frequency components were reconstructed from the complex conjugate.

Once each subband was sign-modulated, the signal was reconstructed by the inverse of the polyphase/FFT analysis which interpolated, frequency shifted, and recombined

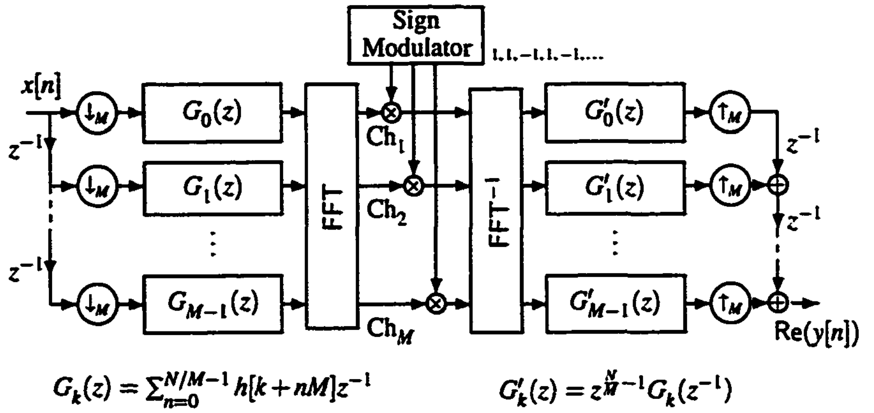


Figure 3.2: Implementation of the polyphase/FFT noise vocoder used in the present experiments. The illustration can be subdivided into three stages: the polyphase/FFT filterbank which decomposes the signal into M critically decimated subbands. The sign of the signal in each subband is then modulated randomly in the second stage. Finally, the subbands are are recombined. Filterbank decomposition and reconstruction was accomplished by a polyphase/FFT realization of a uniform filterbank.

the resulting subbands. This procedure produced stimuli in which the waveform envelope within each channel was preserved while frequency information in each subband was completely distorted.

The examples in Fig. 3.1 on p. 57 were produced in this manner. The center frequency of the lowest and highest channels were at the DC and Nyquist frequencies respectively. Therefore, the cutoff for the lowest frequency filter for the 1000 Hz bandwidth condition was 500 Hz, while the next filter cutoff was at 1500 Hz, *etc.*

Because the filters had linear phase, the temporal distortion was approximately limited to the inverse of the filter bandwidths—*e.g.*, at 500 Hz bandwidth, temporal distortion was effectively 2 ms; at 2000 Hz, temporal distortion was in the order of 0.5 ms. While this may seem large at the narrower bandwidths, it is the smallest amount of temporal distortion that can be expected in any form of noise-vocoded speech. When the channels were recombined without sign-modulation, almost perfect reconstruction was achieved with very little temporal or spectral distortion. Therefore, unintentional consequences of the signal processing were minimized by the present implementation.

3.2.3 Stimulus measurements

Both the gross and detailed spectral measurements required that the burst and formant onsets be located in each of the stimulus waveforms. The initial stop burst was isolated with the aid of a waveform and spectrogram display as was the onset of the vocalic portion. Because we expect very minimal temporal distortion from the signal processing described in Sec. 3.2.2, the location and duration of the burst and formant onsets as measured from the unprocessed stimuli were also used for noise-vocoded stimuli. Otherwise, measurements were obtained from unprocessed and noise-vocoded stimuli independently except where noted below.

Gross spectral shape features were represented by Mel-frequency cepstral coefficients (MFCCs) as described in Sec. 2.2.1. The first 13 MFCCs were measured from a 25 ms frame centered at the onset of the release burst. Another set of 13 MFCCs was measured from a 25 ms frame centered at the onset of the formant frequencies. Both frames were weighted by a 25 ms Hamming function. As can be seen from 2.2 on p. 35, the first cepstral coefficient measures the energy balance between the upper and lower frequency regions and can be interpreted as a measure of gross spectral tilt. The second coefficient measures the energy difference in the mid-frequency region *vs.* the upper and lower frequency regions and may be interpreted as a measure of spectral compactness. Higher coefficients provide additional levels of spectral complexity.

Detailed spectral features were represented by burst peak and formant frequencies. For the burst peak frequencies, the power spectrum was calculated for a single 25 ms window centered at the onset of the burst. This window was also weighted by a 25 ms Hamming function and the resulting spectrum was smoothed using an 16th order LPC model. The frequency of the strongest energy peak above 1 kHz in the LPC-smoothed spectrum was defined as the burst peak frequency. Because the stimuli occasionally had very short VOTs, setting the threshold at 1 kHz ensured that low-frequency voicing was not confused with the burst peak.

Measurement of the formant frequencies was considerably more complicated. The stimuli were downsampled to 8 kHz and pre-emphasized by a factor of 0.97. The extent of voicing was located interactively using a waveform and spectrographic display. The fundamental frequency of the voiced segment was then estimated using an algorithm described by Kawahara, Masuda-Katsuse, and de Cheveigné (1999). With the aid of this initial estimate of f_0, cursors were drawn to roughly separate glottal epochs in the waveform. The spacing between cursors was determined by the estimated glottal period from f_0. However, the position of the first cursor had to be placed manually. At the next stage, there was an opportunity to correct errors in the f_0 tracking by interactively adding, moving, and deleting cursors.

Glottal epochs were represented by the lowpass-filtered magnitude Hilbert transform which gave a rough estimate of the instantaneous energy of the signal (the lowpass-filtered, full-wave-rectified waveform could have been used equally well). Peaks in the magnitude Hilbert transform roughly correspond to the glottal energy peaks in the original signal. The cutoff of the lowpass filter could be adjusted until energy peaks were smooth and regularly spaced. Once the cursors were felt to be well positioned, the algorithm determined energy maxima between each pair of cursors from the lowpass magnitude Hilbert transform. Thus each peak located in such a manner roughly corresponded to a single glottal epoch.

From this, it was possible to perform a pitch-synchronous LPC analysis (Matthews, Miller, and David, 1961). However, based on the suggestion by Smits (1994), each analysis frame spanned exactly two glottal periods and was windowed with a Hamming function. Therefore adjacent analysis frames overlapped by exactly one glottal epoch and the length of each window was adjusted dynamically based on the pitch period. This ensured that spectral distortion caused by small errors in the estimate of glottal periods would be minimized.

Noise-vocoded stimuli were measured slightly differently, however. Because vocal harmonics were no longer present in the noise-vocoded speech, it was felt that an excessively wide bandwidth formant analysis would further reduce the spectral resolution unnecessarily. Therefore, for noise-vocoded stimuli, segments of 25 ms in duration instead of two glottal pitch periods were analyzed.

Linear predictor coefficients for each frame were estimated via the modified covariance method (Kay, 1988). The order of the analysis was determined interactively based on the number of visible formants on a spectrogram—*i.e.*, two coefficients per formant plus one additional coefficient to adjust the overall spectral tilt.

The first three formants were then tracked using a simple dynamic-programming algorithm similar to that proposed by Talkin (1987). A Viterbi algorithm was used (Rabiner, 1989) which minimized the sum of the absolute deviations between pole frequencies in adjacent frames simultaneously for all three formants. Because of the greater between-frame variation in pole frequencies associated with the higher formants, it was always the case that this algorithm followed the first three formants for the 324 syllables used in this stimulus set. Although many other formant tracking techniques have been proposed elsewhere (*e.g.*, McCandless, 1974; Markel and Gray, 1976; Kopec, 1986), this one appeared to be the simplest and most effective as the formants were already well isolated by the LPC analysis.

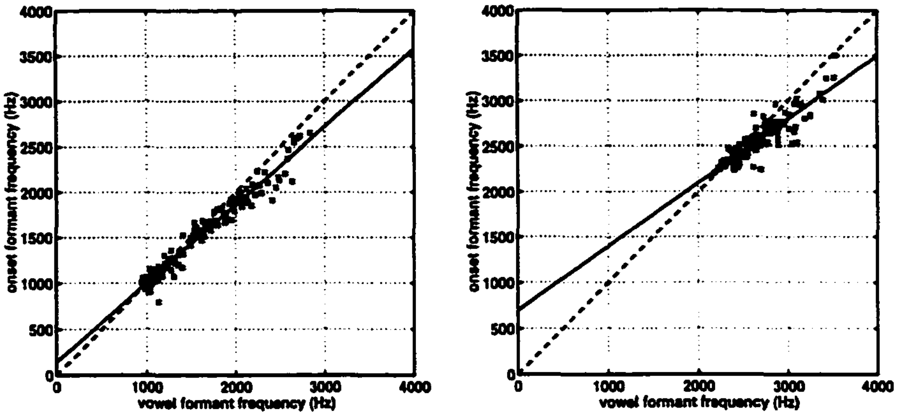
Once the three formant values were estimated for each glottal epoch in this manner, the three formant tracks representing F_1, F_2, and F_3 were finally smoothed using a third order polynomial which allowed for one point of inflection as well as a local minimum and maximum. The polynomial regression was further weighted by the inverse of the estimated bandwidth for each formant at each frame. This ensured that poor formant estimates with a wide bandwidth/low amplitude would not leverage the polynomial fit. The polynomial smooth also gives greater degrees of freedom (larger "hat values") to the endpoints of the formant track where rapid formant movement is normally expected.

As an added check on the quality of the formant tracking, the stimuli were resynthesized using a MATLAB implementation of the cascade branch of the KLATT80 formant synthesizer (Klatt, 1980). The period of the estimated glottal peaks was three-point median smoothed and was used to generate a train of impulses which was then filtered by a single pole at 0 Hz with a -6 dB/octave skirt. The formant bandwidths were fixed at 90 Hz, 110 Hz, 170 Hz, and 400 Hz for F_1–F_4 respectively. The fourth formant frequency was determined by estimating the average spacing between measured formants. F_4 formant frequency was therefore set to the F_3 formant frequency plus the average difference between adjacent formant frequencies for the first three formants. All of the resynthesized stimuli were informally judged by the author as quite satisfactory.

Based on the polynomial-smoothed formant tracks, formant frequencies of F_1–F_3 were measured at the onset of voicing as well as the point 60 ms following this which will be referred to as the *onset* F_{1_o}–F_{3_o} and *vowel* F_{1_v}–F_{3_v} formant frequencies respectively. The detailed spectral measurements are therefore analogous to the ones made by Nearey and Shammass (1987).

Figures 3.3, 3.4, and 3.5 give the F_2 and F_3 locus equations for [b], [d], and [g] respectively as obtained from the measurements described above. The locus equations are represented by the linear regression fits to the vowel and onset formant frequency data and are shown as solid lines in the graphs. The point at which $F_{2_v} = F_{2_o}$ or $F_{3_v} = F_{3_o}$ (indicated by the intersection of the locus equation and the dashed line) gives an estimate of the characteristic locus for each place of articulation and are provided in the captions. Therefore, formant transitions will be rising or falling depending on whether the vowel formant frequency is higher or lower than the estimated locus respectively.

These loci can be compared with previous estimates. For example, Delattre *et al.* (1955) concluded that the second formant loci for bilabials, alveolars and velars were 720 Hz, 1800 Hz, and 3000 Hz respectively. The estimates obtained here are relatively



(a) F_2 locus ≈ 1029 Hz

(b) F_3 locus ≈ 2313 Hz

Figure 3.3: Bilabial locus equations for F_2 and F_3. The locus equation regression fit is indicated by the thick solid line. The intersection of this and the thick dashed line indicates the estimated locus for this place of articulation.

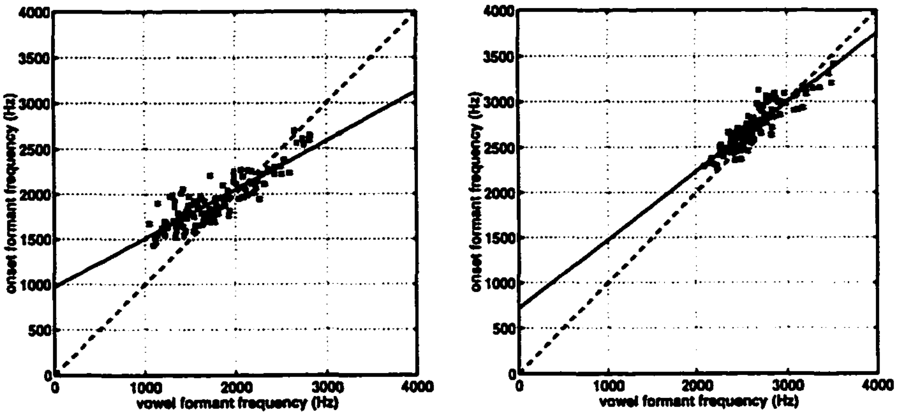
high compared to these values. However, in a similar study, Nearey and Shammass (1987) obtained locus estimates of 1129 Hz, 2084 Hz, and 3150 Hz for [b], [d], and [g] respectively for speakers in the same dialect region. For third formant transitions, Fischer-Jørgensen (1954) obtained estimates of 2300 Hz and 2700–2800 Hz for bilabials and alveolars respectively, which are also very similar to the values found here. Fischer-Jørgensen suggested, however, that velars had no F_3 locus and were generally found to be rising irrespective of the vowel third formant frequency. This is certainly the case here, as F_3 never falls below 100 Hz. Likewise, velar F_2's are never above its characteristic locus, and therefore velar second formant transitions are always falling.

The slope and y-intercepts of the locus equations can be compared with those from similar studies in Table 3.1 on p. 66. In addition, the coefficients of correlation r for each regression is given.

3.2.4 Paired-comparison bootstrap test

The formant tracking procedure described in Sec. 3.2.3 did not prove to be particularly robust in the noise-vocoded stimuli. Unfortunately, most formant tracking algorithms seem somewhat *ad hoc* and assume that formant frequencies have been measured from clean-speech stimuli. Nevertheless, it was felt that the signal processing effects on the formant tracking algorithm could be made distinct from the issue of whether formant frequency information is *potentially* recoverable from noise-vocoded speech by choosing an appropriate nonparametric test.

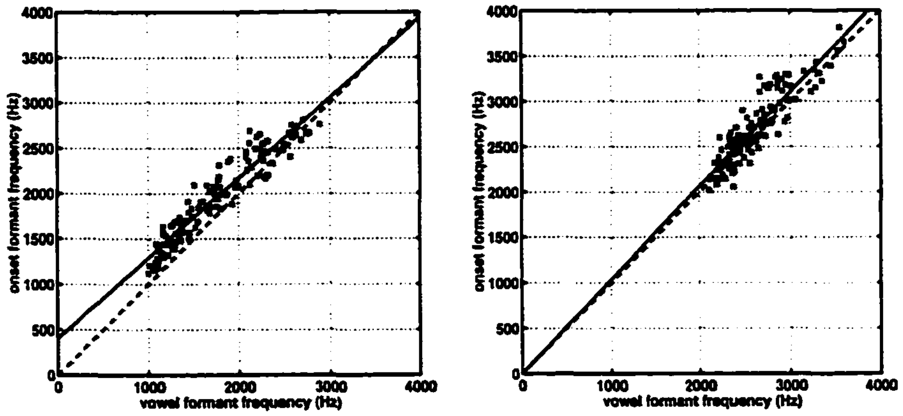
In this vein, a "paired-comparison bootstrap test" was designed which avoided the issue of formant tracking altogether. One randomly selected token A was sampled from the set of unprocessed stimuli. The formant frequencies of A were then compared to those of candidate formants in the corresponding noise-vocoded stimulus A^*; for each of F_1, F_2, and F_3, the corresponding formant frequency was chosen as the pole with



(a) F_2 locus ≈ 2102 Hz

(b) F_3 locus ≈ 2979 Hz

Figure 3.4: Alveolar locus equations for F_2 and F_3



(a) F_2 locus ≈ 3358 Hz

(b) F_3 locus ≈ 100 Hz

Figure 3.5: Velar locus equations for F_2 and F_3

C	Kiefte (2000)	Nearey and Shammass (1987)	Sussman *et al.* (1991)[a]
		second formant	
[b]	F_{2o}=144+0.86F_{2v}; r=0.98	F_{2o}= 192+0.83F_{2v}; r=0.95	F_{2o}= 88+0.89F_{2v}
[d]	F_{2o}=967+0.54F_{2v}; r=0.87	F_{2o}=1042+0.50F_{2v}; r=0.82	F_{2o}=1211+0.42F_{2v}
[g]	F_{2o}=403+0.88F_{2v}; r=0.95	F_{2o}= 215+0.99F_{2v}; r=0.94	F_{2o}= 792+0.71F_{2v}
		third formant	
[b]	F_{3o}=694+0.70F_{3v}; r=0.88	F_{3o}= 945+0.61F_{3v}; r=0.85	...
[d]	F_{3o}=715+0.76F_{3v}; r=0.90	F_{3o}=1344+0.52F_{3v}; r=0.81	...
[g]	F_{3o}= -3+1.03F_{3v}; r=0.90	F_{3o}= 337+0.92F_{3v}; r=0.82	...

Table 3.1: Comparison of locus equations for F_1 and F_2 between present data and that of Nearey and Shammass (1987) and Sussman *et al.* (1991). The coefficient of correlation r is also given.

[a] Sussman *et al.* do not give locus equations for F_3 or coefficients of correlation r for pooled speaker data.

frequency nearest to that of the original stimulus:

$$F^{NV} = f_i^* \ni i = \arg\min_j |f_j^* - F| \qquad (3.1)$$

where F^{NV} is the estimate of the formant frequency from the noise-vocoded signal, f_i^* is the ith pole from the LPC analysis of the noise-vocoded signal, and F is the frequency of the original, unprocessed formant.

This procedure was also repeated for a randomly sampled, unprocessed control stimulus B. If the signal processing distorted formant frequency information to the point where measurements are truly random, we expect noise-vocoded poles to be nearer either to those of the unprocessed stimulus from which it was derived, or to a randomly selected stimulus, with approximately equal probability. Deviations from completely random behavior can then be evaluated for their statistical significance. Significant, nonrandom behavior would indicate, therefore, that some of the original formant frequency information is potentially available in the perception of spectrally reduced stimuli, irrespective of the formant-tracking algorithm used to extract them. Based on this procedure, it is possible to obtain a rough estimate of the *entropy* of noise-vocoded formant measurements.

The procedure is further illustrated in Fig. 3.6 on the next page. 10 000 such iterations were performed for each of four channel bandwidths: 250 Hz, 500 Hz, 1000 Hz, and 2000 Hz. This procedure was performed for MFCCs as well.

3.2.5 Results

Figure 3.7 on p. 68 shows the results of the paired-comparison bootstrap test for the burst peak and formant frequencies. The dashed line indicates the probability that must be exceeded for a significant result at the $\alpha = .05$ level of significance based on the binomial distribution $B(10000, \frac{1}{2})$. As indicated in the figure, formant frequencies are potentially more robust than had previously been anticipated; at 1000 Hz bandwidth,

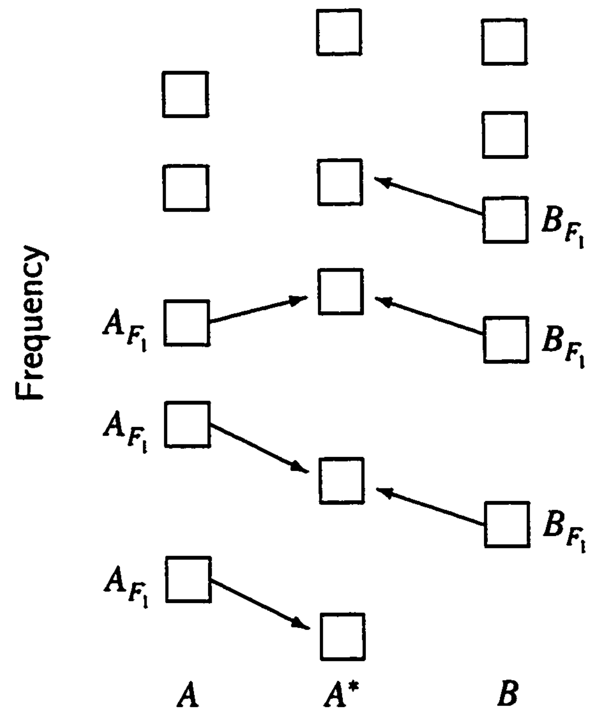


Figure 3.6: Paired-comparison bootstrap test. A and B are randomly sampled with replacement from the full set of unprocessed stimuli, while A^* is the noise-vocoded counterpart to A. In this example, the first three poles of A^* correspond to F_1–F_3 of A, while the second, third and fourth poles of A^* are nearer to F_1–F_3 of B.

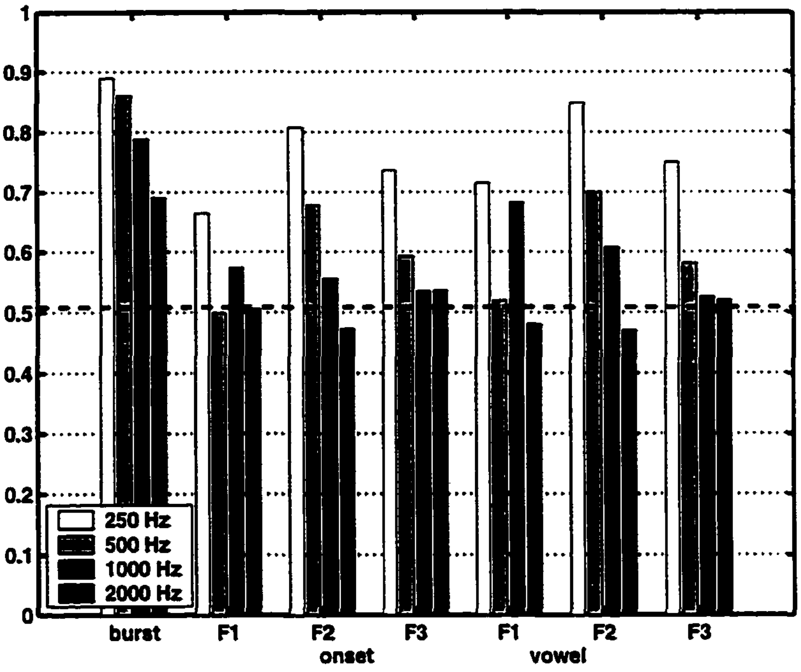


Figure 3.7: The figure shows the results of the paired-comparison bootstrap test for the detailed spectral cues in four vocoder bandwidth conditions, including formant frequencies at the onset of voicing and the vowel as well as the burst peak frequency. The bars show the probability that frequency measurement A_F^* from noise-vocoded stimulus A^* is closer to the corresponding frequency A_F than B_F, where A is the unprocessed stimulus from which A^* is derived and B is a randomly selected control token. The dashed line indicates the probability that must be exceeded for a significant result.

there is still a significant amount of nonrandom formant frequency information present in the noise-vocoded stimuli. The only exceptions to this are the tests for F_1 at the onset of voicing at the 500 Hz bandwidth condition. In general, formant frequencies show significantly nonrandom behavior up to 1000 Hz. At 2000 Hz, however, measurements of F_1 and F_2 appear to be completely random.

In contrast, the burst peak frequency remains nonrandom up through 2000 Hz bandwidth and is potentially a very robust cue in the context of reduced frequency resolution. These results would seem to indicate that the perception of noise-vocoded speech could be performed on the basis of at least some detailed spectral information. Nevertheless there is also a significant decrease in the test statistic between 250 Hz and 500 Hz for all stimulus parameters based on a binomial distribution ($\alpha \ll 0.001$ for all comparisons) indicating that some information is lost at even very small analysis bandwidths.

Figures 3.8 and 3.9 illustrate the results of the paired comparison bootstrap test for the first seven MFCCs for segments centered at the burst and vocalic onsets respectively. With the one exception of the fourth cepstral coefficient at the onset of voicing, the graph shows that all the tests were significantly nonrandom at all vocoder

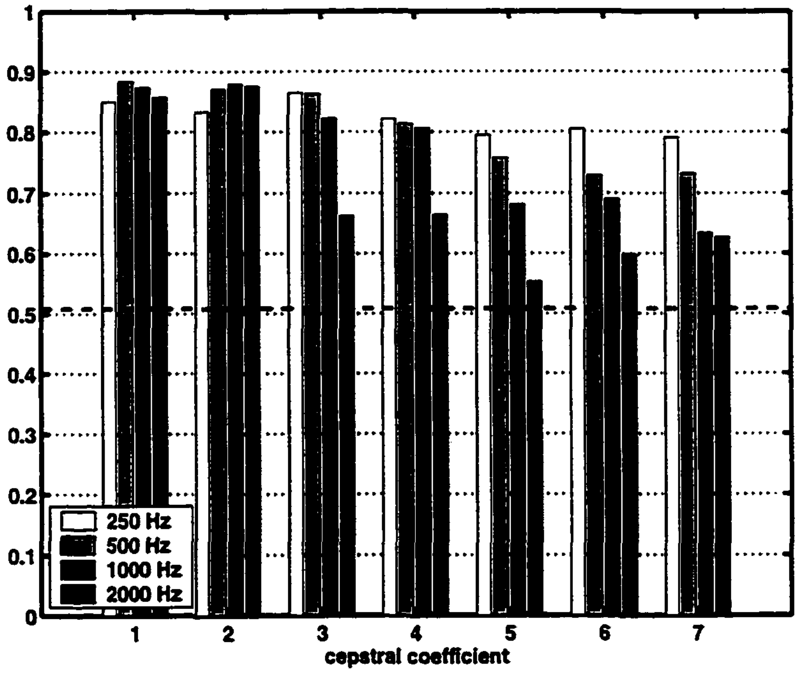


Figure 3.8: Paired comparison bootstrap results: burst gross cues

bandwidths. As expected, the spectral shape parameters are highly nonrandom in the context of noise vocoding.

Although it is clear that formant frequency measurements are nonrandom in noise-vocoded stimuli up to 1 kHz bandwidth, these tests cannot tell us the *magnitude* of the distortion. Figure 3.10 on p. 71 is an attempt to illustrate this. The figure shows the root-mean-squared (rms) errors between measurements from unprocessed and corresponding noise-vocoded stimuli:

$$\mathrm{rms} = \left[\frac{1}{N}\sum_{i=1}^{N}(F_i^{NV} - F_i)^2\right]^{\frac{1}{2}} \tag{3.2}$$

where F_i is the formant frequency of observation i, F_i^{NV} is the nearest pole frequency in the noise-vocoded counterpart (Eq. 3.1 on p. 66), and N is the number of iterations in the resampling scheme. For comparison, the rms errors found between unrelated stimuli are also included. Although both of these values are biased downwards because of the procedure used for selecting formants from among candidates in the distorted tokens, the *sign* of the difference is expected to be accurate. In addition, the relative differences between paired and random rms values can be somewhat informative. The significance of the difference between these rms errors can be evaluated by comparing the ratio of their squares with an F-distribution with 9999 degrees of freedom in both the numerator and denominator under the null-hypothesis H_0. For a level of significance of $\alpha < .01$, this ratio must be greater than 1.05 and the ratio of the rms values must therefore be greater than $\sqrt{1.05} \approx 1.02$—*i.e.*, all differences are significant un-

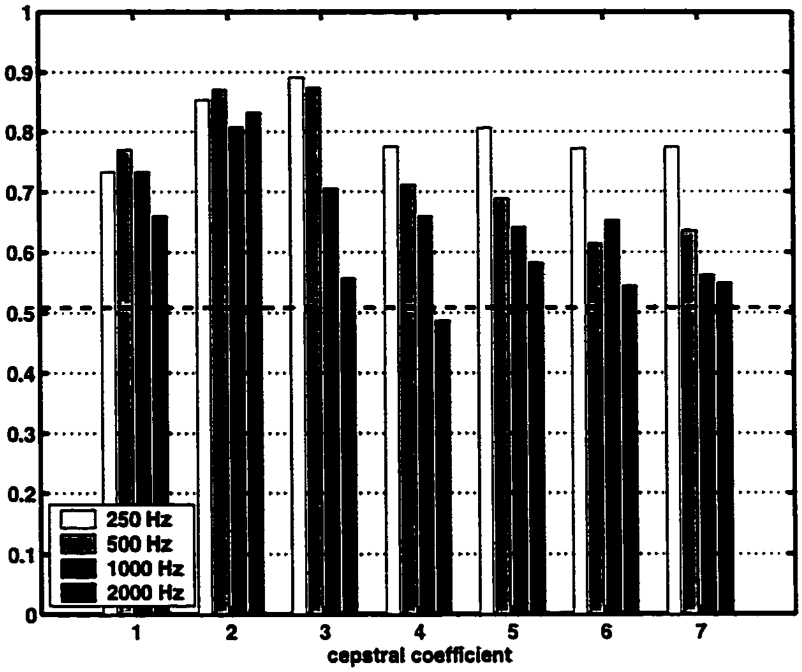


Figure 3.9: Paired comparison bootstrap results: formant onset gross cues

less the plot markers actually overlap. With the exception of F_{1_o} at 500 Hz, it is clear that rms errors for paired stimuli are smaller than those for the control stimuli up to 1000 Hz.

It should be noted that, despite the potential for the preservation of some formant frequency information, it has still not been established *how* this information might be extracted. Unfortunately, as suggested above, the formant tracking algorithm will likely fail for noise-vocoded stimuli; like most other formant tracking algorithms (*e.g.*, McCandless, 1974; Markel and Gray, 1976; Talkin, 1987), it was designed for use with clean-speech stimuli. Although LPC analysis is not the only procedure that has been used to measure formant frequencies, most such algorithms rely on some type of formant tracking to identify specific formants (*e.g.*, Potamianos and Maragos, 1996).

In conclusion, it would appear that MFCCs, which represent gross spectral shape properties, are potentially very robust in the context of reduced spectral resolution. However, some formant and burst peak frequency information also appears to be present up to vocoder bandwidths of 1 kHz. It is entirely possible then that listeners in previous experiments may be relying on some detailed spectral information. Formant frequencies cannot be ruled out as a potential acoustic correlate to place of articulation perception in noise-vocoded speech.

3.2.6 Automatic classification

The purpose of this section is to evaluate the usefulness of estimated gross and detailed acoustic features in the automatic classification of both unprocessed and noise-vocoded

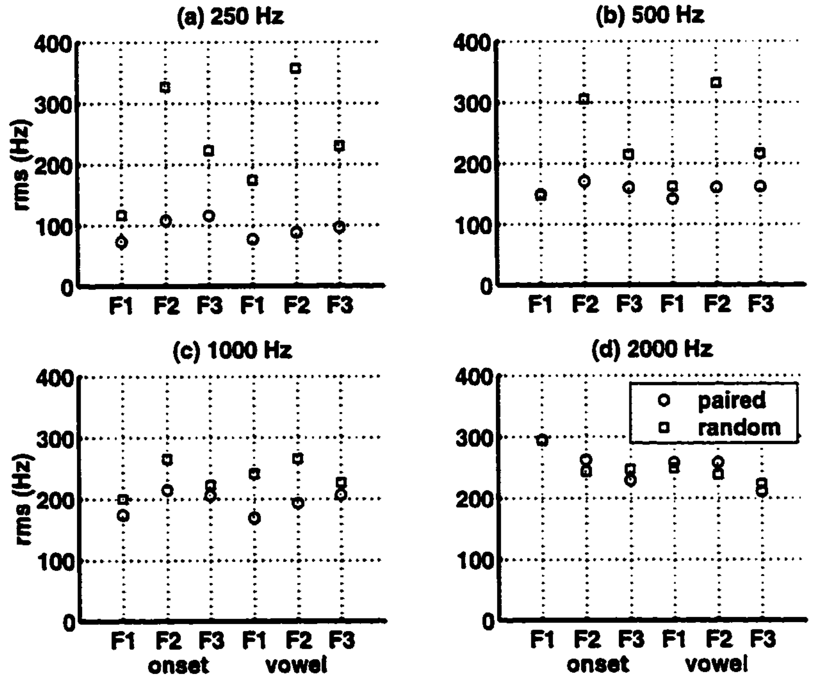


Figure 3.10: Root-mean-squared errors between paired and random control stimulus parameters for each of the formant frequencies and at each vocoder bandwidth.

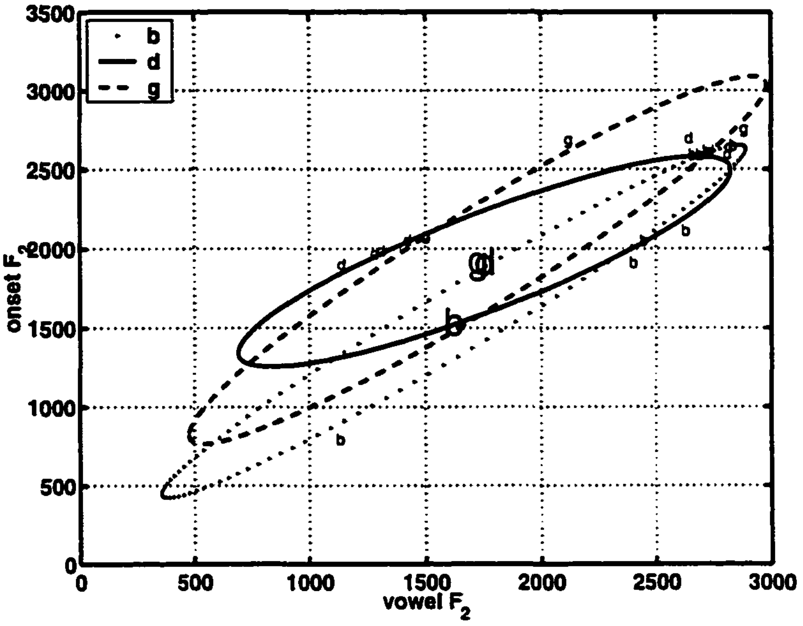


Figure 3.11: Estimated 95% confidence regions for [b], [d], and [g] for F_2 based on a bivariate normal distribution. The shape and orientation of each of the ellipsoids reflects the covariance structure. Outliers—*i.e.*, those outside of the 95% confidence regions—are also indicated.

stimuli. Although it was shown in the previous section that some formant and MFCC measurements were nonrandom in spectrally degraded speech, and that the magnitude of the distortion was relatively small up to 1 kHz vocoder bandwidth, it did not establish how this distortion might affect classification performance. It may be the case that the increase in within-category variance may have a substantial effect on listeners' ability to discriminate the three places of articulation. To address this issue, a quadratic discriminant analysis (QDA) was performed based on the extracted gross and detailed features from unprocessed stimuli. Parameters estimated from the QDA were then used to classify noise-vocoded stimuli.

The 95% confidence regions based on bivariate normal distributions $N_2\{\mu_i, \Sigma_i\}$ for (F_{2_v}, F_{2_o}) and (F_{3_v}, F_{3_o}) are represented in Figs. 3.11 and 3.12 respectively as three overlapping ellipsoids for each place of articulation. The ellipsoid centers are located at the estimated means $\hat{\mu}_i$ for each place of articulation i and their shapes and orientations are determined by the estimated covariance matrices $\hat{\Sigma}_i$. The Cholesky decomposition $\hat{\Sigma}_i^{1/2}$ was then scaled by $(\chi^2_{2,.95})^{1/2}$ to produce the ellipsoids. Individual observations that fall outside these confidence regions are also indicated.

This type of graphical representation is a much better illustration of the potential discriminability of the three categories than the raw locus equations themselves. As can be seen from the figure, the means for [d] and [g] are almost identical for both F_2 and F_3 indicating that the only distinguishing characteristic is the covariance between the onset and vowel measurements. This suggests that a linear discriminant analysis—

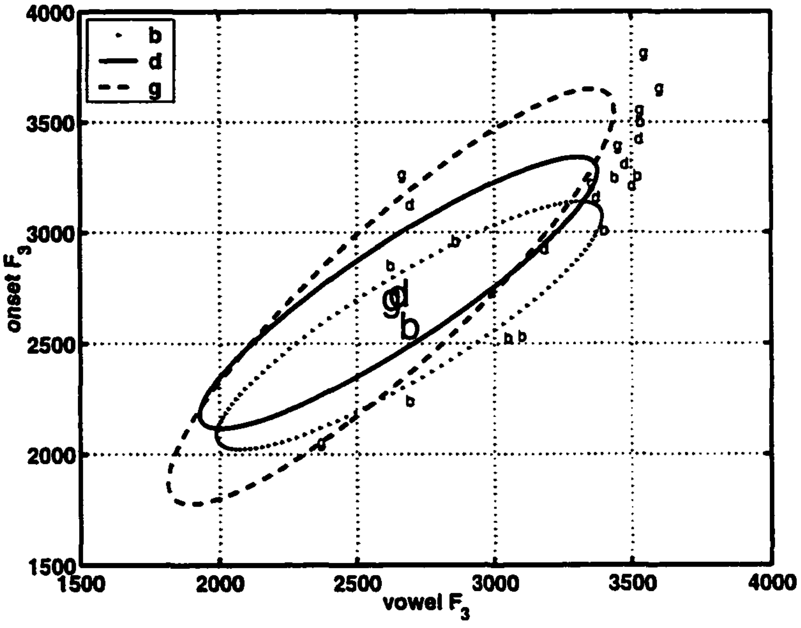


Figure 3.12: Estimated 95% confidence regions for [b], [d], and [g]: F_3

which assumes that all categories have a common covariance matrix $\hat{\Sigma}$— would be inappropriate for this data.

For both of the detailed and gross feature sets, a leave-one-speaker-out cross-validation was performed on the unprocessed stimuli via QDA. Means and covariance matrices estimated from the QDA were then used to classify noise-vocoded stimuli—*i.e.*, QDA was performed on the unprocessed syllables of 11 speakers and was used to classify the ***noise-vocoded*** stimuli of the 12th remaining speaker.

However, for the detailed spectral features, the procedure outlined in Eq. 3.1 on p. 66 was retained. This has the potential to bias the classification results of noise-vocoded stimuli towards better performance. Nevertheless, it is a good estimate of the ***upper bound*** of potential discriminability given a formant tracker optimized for spectrally degraded speech, even if no such formant tracker exists.

Figure 3.13 on the following page shows the results of this test. As shown in the figure, LPC coefficients achieve significantly better classification rates than up to nine MFCCs for unprocessed speech. However, classification of 1000 Hz noise-vocoded stimuli using formant and burst peak frequencies obtains lower classification rates than just three MFCCs. For the gross spectral features, more than three cepstral coefficients offers little no significant improvement in the 1 kHz bandwidth condition, while the additional parameters actually worsens correct classification performance at the 2 kHz condition. This suggests that the reliance on more complex detailed spectral shape parameters is not robust in the context of spectral distortion. It is also interesting to note that no difference in correct classification performance is observed for MFCCs between unprocessed, 250 Hz, and 500 Hz noise-vocoded stimuli using up to nine coefficients. In contrast, detailed spectral measures rapidly degrade even at the smallest levels of

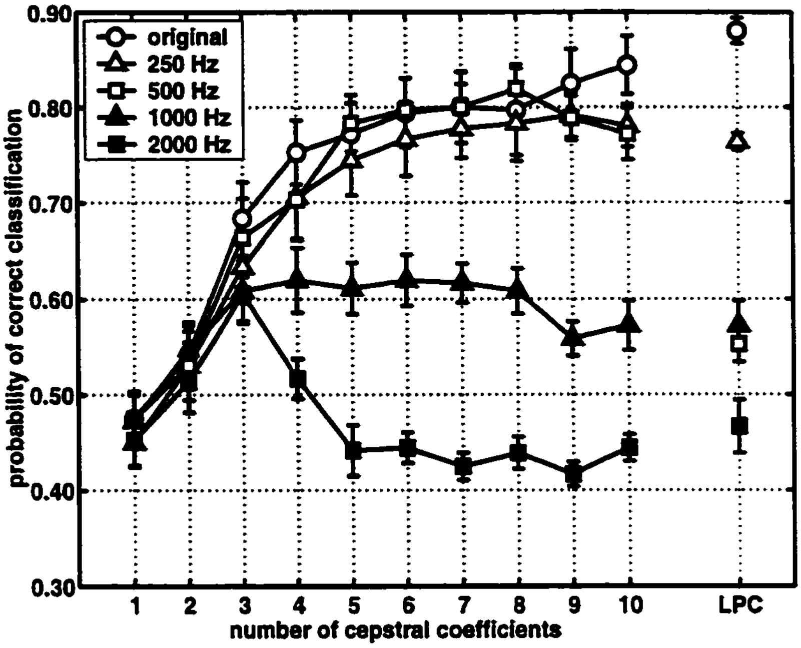


Figure 3.13: Leave-one-speaker-out cross-validation probability of correct classification using gross spectral shape properties (MFCCs) and detailed formant and burst peak frequency measurements (LPC) based on quadratic discriminant analysis. The error bars indicate ± one standard error.

distortion and the decrease in performance at 500 Hz is relatively large in magnitude.

Two things should be highlighted in these results: although the performance of the LPC model is biased upwards in the case of noise-vocoded speech, it is expected to be unbiased for unprocessed speech, since an explicit formant tracker was used. Therefore, on the basis of the performance on clean-speech stimuli, detailed measures are generally superior to the gross spectral features. Although this conflicts with observations made by Nossair and Zahorian (1991), who found that discrete cosine coefficients performed consistently better than formant frequencies in an automatic classification task involving prevocalic stop consonants, fewer data are used here and the complexity of the gross spectral shape model is limited by the available degrees of freedom. Nevertheless, there is a sharp drop in correct classification performance for the LPC model for even the smallest levels of spectral distortion, despite the fact that this performance is overestimated. This indicates that classification of noise-vocoded stimuli on the basis of detailed spectral cues in the form of burst and formant frequencies is expected to do much more poorly than the gross spectral cues represented by Mel-frequency cepstral coefficients.

Based on these results, we might expect that if listeners depend largely on detailed spectral properties, correct identification performance in a perceptual task should decrease significantly at 500 Hz vocoding bandwidth. Conversely, if listeners attend mostly to global spectral shape properties, they should not show significant decreases in performance except for 1000 Hz or wider bandwidth noise-vocoded stimuli. This would, in fact, correspond to the definition of detailed *vs.* gross spectral features as defined by Smits *et al.* (1996b) who suggested that detailed cues are those which require a spectral resolution narrower than 500 Hz. Nevertheless, even under the detailed-cue theory of stop perception, we should not be surprised to see categorization performance exceed chance levels of correct classification, even for 1000 Hz bandwidth noise-vocoded stimuli.

Alternatively, listeners may only use the least detailed spectral features, such as overall spectral slope and compactness, in which case, they may show no significant detriment to performance for either of the 500 Hz or 1000 Hz bandwidth conditions. For example, this is roughly the result obtained by Dorman *et al.* (1997) who found that perception did not improve for more than six vocoder channels, whose bandwidths are close to 1000 Hz in the region of F_2.

Direct comparisons between these predictions and results found in previous experiments are difficult to make, as unequal bandwidths were used in the studies by Shannon *et al.* (1995) and Dorman *et al.* (1997). In addition, it has already been noted that subjects in these experiments received extensive training. Therefore, the next experiment evaluates listeners' ability to perceive place of articulation in the uniform-bandwidth noise-vocoded stimuli described in this section. This is an attempt to not only replicate previous experimental results using untrained listeners, but also to compare subjects' correct identification rates with the predictions made here for gross and detailed feature sets.

3.3 Perception experiment

The purpose of this experiment was to evaluate subjects' ability to identify stop-consonant place of articulation from noise-vocoded speech. Although, similar experiments have been described previously (Shannon *et al.*, 1995; Dorman *et al.*, 1997), this experiment was performed to make an explicit comparison between place perception in spectrally degraded speech and the results obtained in the automatic classification experiment in Sec. 3.2.6.

3.3.1 Stimuli

The stimuli used in this experiment were the same as those used in Sec. 3.2 with the exception that both voiced and voiceless stops were present in the stimulus set and only the vowels [e], [æ], [ɒ], and [o] were included. A total of 6 × 4 × 12 = 288 syllables were used in this experiment. In order to evaluate the predictions made in the previous section, all stimuli were noise-vocoded at 500 Hz and 1000 Hz. In addition, stimuli which were processed by the noise vocoder without sign-modulation of the subbands—*i.e.*, reconstructed stimuli—were included as an added check on any unintended distortion caused by the polyphase/FFT decomposition/reconstruction itself.

3.3.2 Subjects

Nine graduate and undergraduate students of Linguistics were paid as subjects in the experiment. None reported any hearing impairment and all were native speakers of Western Canadian English.

	response			
place	lab.	alv.	vel.	pcc
	no noise vocoding			
labial	0.98	0.01	0.00	0.97
alveolar	0.03	0.97	0.01	
velar	0.00	0.03	0.96	
	bandwidth = 500 Hz			
labial	0.91	0.07	0.03	0.89
alveolar	0.09	0.88	0.04	
velar	0.03	0.09	0.88	
	bandwidth = 1000 Hz			
labial	0.87	0.06	0.06	0.76
alveolar	0.17	0.80	0.03	
velar	0.16	0.24	0.60	

Table 3.2: Confusion matrices for reconstructed and noise vocoded-stimuli. Rows give the original place of articulation while columns give the response probabilities. Values under the column "pcc" give the overall probability of correct classification.

3.3.3 Procedure

Stimuli were presented to subjects for identification of the syllable-initial stop. Stimuli were completely randomized and presented in two sessions of approximately 25 minutes each. In total, 3 x 288 = 864 responses were given by each subject. Subjects indicated responses by clicking on the appropriate box on a computer display. Listeners received no training prior to listening to the stimuli.

Stimuli were resampled at 44.1 kHz and played through a Gina AD/DA at 16 bit quantization on a PC. Subjects listened to stimuli over a loud-speaker in a sound-treated room at a comfortable listing level.

3.3.4 Results

Table 3.2 gives pooled confusion matrices for stop place of articulation identification. The stimulus place is given down the columns while the response is given along the rows. As illustrated in the table, there was a decrease in correct identification performance for the noise-vocoded stimuli. All of these values are significantly better than chance ($\alpha \ll 0.001$) based on a binomial distribution $\mathrm{B}(N, \frac{1}{3})$ where N is the total number of responses given by all subjects. The most noticeable effect is on the identification of velars which drops to 60% correct in the 1000 Hz bandwidth condition.

In order to evaluate the differences between the three conditions, McNemar test statistics (Eq. 2.2 on p. 43) were calculated for each subject (Fleiss, 1981). These are described in more detail in Sec. 2.3.5. The McNemar test statistic for each subject can be compared to a χ_1^2 distribution under the null hypothesis H_0. However, a more exact test refers n_A to a binomial distribution $B(n_A + n_B, \frac{1}{2})$ where n_A and n_B are the number of errors made in condition A not made in condition B and *vice versa* (Ripley, 1996). Table 3.3 on the next page shows the significance of this latter comparison under the

	recon.			500 Hz			1000 Hz
Subj.	%	m	α	%	m	α	%
A	98	2.29	0.06	94	16.53	0.00^{**}	78
B	92	4.65	0.01^{*}	83	5.30	0.01^{*}	73
C	98	1.13	0.14	95	11.25	0.00^{**}	84
D	98	4.90	0.01^{*}	92	4.05	0.02^{*}	85
E	98	3.27	0.03^{*}	93	9.03	0.00^{**}	81
F	99	1.78	0.09	95	12.90	0.00^{**}	81
G	99	6.13	0.00^{**}	93	21.19	0.00^{**}	73
H	97	4.50	0.02^{*}	90	19.53	0.00^{**}	72
I	100	5.14	0.01^{*}	95	12.00	0.00^{**}	82

Table 3.3: Correct place of articulation identification scores for each subject. Percent correct identification is given under the columns labelled "%" for the three processing conditions used in this experiment. See Table 2.1 on p. 44 for an explanation of the columns m and α. These statistics are based on the changes in identification between each noise-vocoding condition.
$^{*}\alpha < 0.05$; $^{**}\alpha < 0.01$

columns labeled α. The differences between the reconstructed and 500 Hz bandwidth conditions are significant for all but three of the subjects ($\alpha < 0.05$). The differences between the 500 Hz and 1000 Hz bandwidth conditions are highly significant for seven of the nine subjects ($\alpha < 0.01$) and significant for the remaining two subjects ($\alpha < 0.05$). The sum of the McNemar test statistics m can be compared to a χ^2_N distribution under the null hypothesis where N is the number of subjects. The differences between the unprocessed and 500 Hz bandwidth condition, as well as the differences between the 500 Hz and 1000 Hz conditions were highly significant ($\chi^2_9 = 33.78$; $\alpha \ll 0.001$ and $\chi^2_9 = 111.78$; $\alpha \ll 0.001$ respectively).

Based on the predictions made in Sec. 3.2.6, the significant decrease in correct identification performance at 500 Hz bandwidth suggests that listeners may be relying on detailed spectral properties. However, the *magnitude* of the decrease is perhaps much less than would be predicted by the results in the automatic classification task.

3.3.5 Modeling of listeners' responses

Using the cross-validation procedure described in Sec. 2.3.5, listeners responses to both unprocessed and noise-vocoded stimuli were modeled via a multinomial loglinear regression using either detailed or gross cues as variates. This procedure operates under the assumption that the underlying distribution of the stimulus parameters is approximately multivariate normal $N_p\{\mu_i,\Sigma\}$. This has been referred to as the normal *a posteriori* probability model of phoneme recognition (Nearey and Assmann, 1986; Andruski and Nearey, 1992; Hillenbrand and Nearey, 1999). However, by default, the model also assumes that the covariance between continuous variates Σ is equal for all catogories. This is clearly not a good choice based on the relations illustrated in Figs. 3.11 and 3.12. In order to capture the observed covariation between detailed spectral parameters, it was decided to also include the squares of all the variates, as well as interaction effects between vowel formant frequencies and the corresponding onset formant frequencies. In addition, as it has been observed that the perceived category of the release burst depends on the vocalic context in which it occurs (Liberman *et al.*, 1952; Schatz, 1954), the interaction effect between the burst peak and F_{2v} vowel formant frequency was also included.

Although the automatic classification procedure described in Sec. 3.2.6 assumed that the covariance matrices were completely independent for the three categories Σ_i, an analysis of these response data based on this assumption would leave very few residual degrees of freedom. It was therefore decided to use a more conservative representation that focused on the covariance relations that have been observed in previous experiments—such as the relationship between the onset and vowel formant frequencies of F_2 and F_3. In addition, F_1 was completely omitted from the analysis as it has not been shown to carry a significant amount of place of articulation information (Delattre *et al.*, 1955). Therefore, the model used here assumes that the onset and vowel formant frequencies covary independently between different places of articulation, but that the covariation between frequencies *across* formants is held fixed for all categories.

In total, 13 parameters were used: F_{burst}, F_{2o}, F_{3o}, F_{2v}, and F_{3v}; the squared variates F^2_{burst}, F^2_{2o}, F^2_{3o}, F^2_{2v}, and F^2_{3v}; and the interactions $F_{\text{burst}} \times F_{2v}$, $F_{2o} \times F_{2v}$, and $F_{3o} \times F_{3v}$.

Other modeling experiments have used the orthonormal distance between individual observations to the locus-equation regression lines as illustrated in Figs. 3.3, 3.4, and 3.5, in estimating category membership for each observation (*e.g.*, Nearey and Shammass, 1987; Smits *et al.*, 1996b). Conceptually, this representation can be compared to a transformation of the covariance ellipsoids in which the first principal component is infinitely long while the second principal component is equal for all categories—*i.e.*, the regression lines represent infinitely long ellipsoids with equal width. In this case, the length of the secant vector parallel to the principal axis of the covariance ellipsoid is undefined and the distance to the category center is determined entirely by the orthonormal distance to the principal axis. Effectively, this model ignores the distribution of observations along the length of the locus equation regression line. Nevertheless, this relationship is expected to be sufficiently well represented by the "conservative" quadratic loglinear regression described above, provided that the models are not overfit.

In order to provide an unbiased estimate of noise-vocoded formant frequencies, the formant tracker implemented in the speech analysis program PRAAT was used. This is a dynamic-programming algorithm similar to the one described by Talkin (1987) which minimizes absolute differences in formant frequences between adjacent frames, formant bandwidths, and the total deviation from the "average" formant values of 550 Hz, 1650 Hz, and 2750 Hz for F_1, F_2, and F_3 respectively. Five poles between 0 Hz and 5500 Hz were measured via the Burg method of autoregressive parameter estimation in 25 ms frames with an overlap rate of 15 ms. Once formant identities were established by the dynamic-programming formant-tracking algorithm in PRAAT, frequency values were then smoothed by a third order polynomial function as described in Sec. 3.2.3 to reduce the variance of the formant measurements. Formant frequencies were then taken from the smoothed formant estimates.

Figure 3.14 on the following page gives scatter plots of the estimates of F_2 formant frequency for the three processing conditions. The distribution of observations in the 500 Hz bandwidth condition appears to be relatively conservative when compared to that found in the 1000 Hz bandwidth condition.

The predictive performance of the detailed spectral model was then compared to that of the gross spectral cue model represented by MFCCs measured from 25 ms

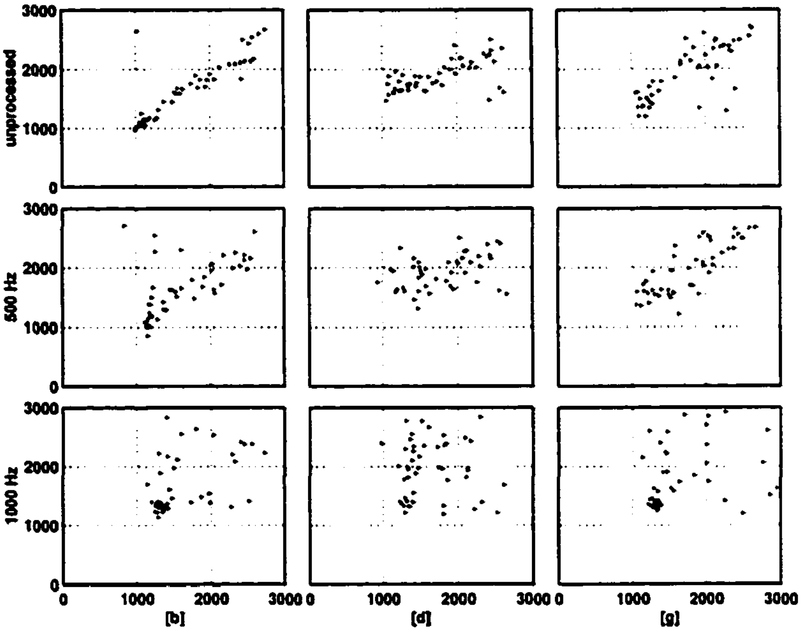


Figure 3.14: F_{2v} by F_{2o} scatter plots from estimated formant frequencies. The x-axis indicates the F_2 formant frequency at the vowel, while the y-axis represents the F_2 formant frequency at the onset of voiced formants.

	1 frame		2 frames		3 frames	
#	pma	rms	pma	rms	pma	rms
1	0.00	0.00	0.00	0.03	0.00	0.03
2	0.04	0.35	0.62	0.77	0.66	0.77
3	0.77	0.94	0.98	1.00	1.00	1.00
4	0.93	0.98	1.00	1.00	1.00	1.00
5	0.94	0.98	1.00	1.00	1.00	1.00
6	0.93	0.97	1.00	1.00	1.00	0.99
7	0.91	0.96	1.00	1.00	0.99	0.99
8	0.94	0.99	1.00	1.00	0.99	0.99
9	0.92	0.98	1.00	1.00	0.99	0.98
10	0.91	0.96	1.00	1.00	0.99	0.98

Table 3.4: Cross-validation comparisons between gross and detailed-spectral cue models. Values give the proportion of trials in which MFCCs obtained better measures of goodness of fit—*i.e.*, larger percentage modal agreement or smaller root-mean-squared error.

Hamming-weighted windows centered at the onset of the burst release, the onset of voiced formant transitions, and the vocalic portion 60 ms folowing the onset of voicing. Measurements were made as described in Sec. 3.2.3. Models consisting of between one and three of these frame segments were considered. In the one-frame model, only the MFCCs from the release burst are included, for the two-frame model, MFCCs from the burst and onset of voicing are included, and all three frames are included in the three-frame model. Between one and ten cepstral coefficients measured from each frame were used in the generalized linear model.

The cross-validation procedure described in Sec. 2.3.6 was performed in a total of 500 iterations to determine the number of trials in which the MFCC models produced better goodness-of-fit measures than the detailed model described above (see Eq. 2.5 on p. 48). At each iteration, the data were partitioned in two stages, the data from four subjects to stimuli from six speakers were used in the training set S_{tt}. The parameters of the model were then held fixed, and were used to predict responses to the remaining stimuli. These predictions were then evaluated on the basis of the remaining data in the assessment set S_{aa} which consisted of the responses from five subjects to the stimuli produced by six speakers. Two measures of goodness-of-fit were considered: percentage modal agreement pma and root-mean-squared error rms (see Sec. 2.3.6).

The results of this test are presented in Table 3.4. The proportion of iterations in which each of the MFCC models' goodness-of-fit scores were better than that of the detailed cue model are given. In the case of percentage modal agreement, improvement in fit is indicated by a larger pma, while the root-mean-squared error is expected to decrease for models that are closer to the actual model underlying listeners' responses and. In both cases, a value greater than 0.95 under the columns labeled pma and rms indicates significantly better performance for MFCCs at the $\alpha = 0.05$ level of significance.

The table shows that with only one cepstral coefficient measured in either one, two, or three segments, the pma and rms scores for detailed spectral cues are significantly better than those of the gross cue model—*i.e.*, proportions are less than 0.05. However,

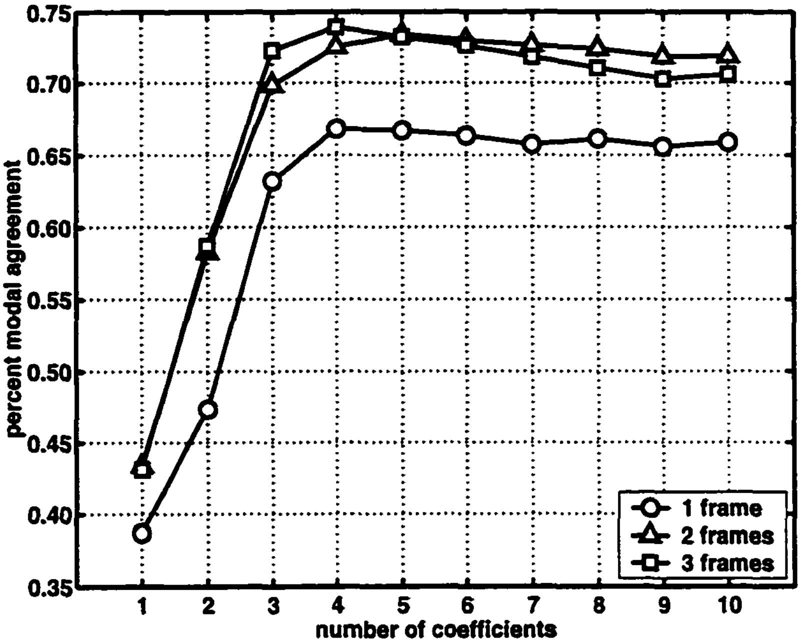


Figure 3.15: Percent modal agreement for MFCC models. The lines indicates the cross-validation estimate of the pma for MFCC models ranging from one to three frames and one to ten coefficients for each frame.

analyses based on three or more cepstral coefficients perform significantly better than the detailed cue model. In addition, some of these models have many fewer degrees of freedom (10 in the case of the MFCC model with four coefficients measured at one frame at the burst onset *vs.* 28 for the detailed-cue model). It would appear then, that only a few Mel-frequency cepstral coefficients are able to predict listeners' responses to noise-vocoded stimuli significantly better than the detailed spectral measures described here.

The cross-validated estimate of percent modal agreement for each of the MFCC models is illustrated graphically in Fig. 3.15. Clearly the models consisting of more than one frame perform better. However, there does not appear to be an appreciable improvement with more than four coefficients indicating that only very broad spectral shape features are necessary to model listeners' responses to these stimuli. These values can be compared with the mean cross-validated pma obtained for the burst peak and formant frequency model which was found to be 59.4%.

It could be argued that the single peak representation of the release burst is relatively impoverished. Other spectral measures have been considered such as the frequencies of the two most promenent peaks (Jongman and Miller, 1991) and the burst amplitude (Ohde and Stevens, 1983; Smits *et al.*, 1996b). In addition, it was shown in Chap. 2 that a *static* representation of the global spectral properties of the release burst was able to model listeners' responses to short gated stops relatively well. It is indeed possible that the release burst is best represented by gross-spectral-shape parameters, while detailed frequency measures are more appropriate for the vocalic portion. There-

	2 frames		3 frames	
#	pma	rms	pma	rms
1	0.00	0.00	0.00	0.00
2	0.09	0.24	0.11	0.28
3	0.63	0.78	0.81	0.89
4	0.84	0.93	0.93	0.95
5	0.92	0.96	0.89	0.92
6	0.88	0.93	0.87	0.86
7	0.88	0.93	0.82	0.81
8	0.88	0.92	0.74	0.69
9	0.83	0.87	0.67	0.53
10	0.84	0.87	0.69	0.54

Table 3.5: Cross-validiation comparisons: hybrid model

fore, a larger *hybrid* model was designed in which the release burst was represented by 10 MFCCs and the vocalic portion was represented by the squares and cross-products of corresponding formant frequencies as described above. This model was similarly compared to the full MFCC models.

Cross-validation was performed in 500 iterations to determine the significance in the difference between the gross and hybrid-spectral-cue models. Table 3.5, which is analogous to Table 3.4 shows the results of this test. The hybrid model was not compared to the one-frame MFCC model which is a subset of the hybrid variates. By comparison with the conservative detailed spectral model above, the differences between the hybrid and full MFCC models appears to be smaller in terms of their ability to predict responses to new stimuli. Only marginally significant improvements were found for the full MFCC model for five coefficients in two frames and four coefficients in three frames. The overall mean pcc score for the hybrid model was 68.3%, which is quite a bit higher than that found for the conservative detailed spectral model (59.4%).

In Sec. 3.2, concerns were expressed with regards to the performance of automatic formant trackers in spectrally degraded speech. In order to make the comparison between MFCC models and the detailed spectral cue model fair, it was decided to use an explicit automatic formant tracker anyway, under the assumption that errors made by this algorithm may be correlated with errors made by human listeners under the same signal conditions. However, this may not be the case. It is possible that subjects' ability to identify formants in spectrally degraded speech is more robust than would be predicted by an *ad hoc* formant tracker designed for the analysis of clean-speech stimuli.

In order to test this hypothesis, the formant tracking algorithm was again circumvented altogether, and formant frequencies in noise vocoded speech were estimated in a manner similar to that described in Sec. 3.2.6. Namely, noise-vocoded formant candidates were selected on the basis of their proximity to their original, unprocessed counterparts. However, in contrast to the formula provided in Eq. 3.1 on p. 66, poles were not assigned to more than one formant. Thus all possible assignments to F_2 and F_3 were considered and the combination that minimized the sum of the absolute deviations from the unprocessed measurements were used.

Cross-validation was again performed in 500 iterations to determine if the MFCC

	1 frame		2 frames		3 frames	
#	pma	rms	pma	rms	pma	rms
1	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.11	0.67	0.14	0.33
3	0.12	0.29	0.75	0.91	0.91	0.95
4	0.44	0.52	0.92	0.97	0.96	0.98
5	0.42	0.51	0.94	0.96	0.95	0.95
6	0.38	0.47	0.92	0.95	0.93	0.92
7	0.32	0.46	0.92	0.95	0.91	0.88
8	0.42	0.62	0.91	0.93	0.86	0.76
9	0.36	0.55	0.89	0.88	0.77	0.64
10	0.38	0.53	0.89	0.89	0.83	0.66

Table 3.6: Cross-validation comparisons of goodness of fit: biased detailed model

models are still significantly closer to the true model underlying listeners's responses than the detailed-spectral model in which formant identities were determined from an external source. The results are shown in Table 3.6. The biased detailed model performs much better against the one-frame MFCC models. However, the three-frame model with four MFCCs still performs significantly better with even fewer degrees of freedom (26 *vs.* 28 for the detailed-cue model). This is despite the fact that the optimized formant frequency model is ***biased*** towards more robust correct classification behavior as in the automatic classification experiment described in Sec. 3.2.6. The estimated mean pma for the biased detailed model was 66.9%, which is quite a bit better than the unbiased model. However, it is not clear if listeners would even be able to track formants with the robustness proposed by this model.

The conclusions drawn in this section assume that the detailed parameters estimated from noise-vocoded stimuli are an accurate representation of perceptual correlates to place of articulation. However, this may not be the case. Two *ad hoc* procedures were implemented to identify candidate formants from noise-vocoded stimuli: one which used an explicit formant tracker designed for the analysis of clean-speech stimuli and another method that biased model performance towards more robust behavior.

3.3.6 Discussion

Table 3.7 on the next page shows that there may be a difference in the way velars are perceived in different vowel contexts. Errors in identification for velar stops tend towards different places of articulation depending on whether the medial vowel is front ([e]) or back ([o])—*i.e.*, there is a stronger bias towards alveolar responses before [e]. In contrast, there is a stronger bias towards bilabial responses before [o]. One possible explanation for this may relate to the robustness of the burst cue relative to that of the formant transitions. It has been found that velar F_2 formant transitions fall by roughly 200 Hz towards the following vowel target (Sussman *et al.*, 1991) which is much smaller than the bandwidth of the narrowest analysis filters used in this experiment. In the absence of a detectible falling F_2 formant frequency, it is possible that a high frequency velar burst characteristic of a front vowel context will cue an alveolar place of articulation, while a low-frequency burst characteristic of a back vowel con-

	response							
	medial vowel: [e]				medial vowel: [o]			
place	lab.	alv.	vel.	pcc	lab.	alv.	vel.	pcc
				No Noise Vocoding				
labial	0.96	0.04	0.00	0.95	0.99	0.00	0.01	1.00
alveolar	0.06	0.93	0.02		0.00	1.00	0.00	
velar	0.00	0.04	0.96		0.00	0.00	1.00	
				Bandwidth = 500 Hz				
labial	0.85	0.08	0.07	0.85	0.97	0.03	0.01	0.95
alveolar	0.12	0.83	0.05		0.05	0.92	0.04	
velar	0.04	0.10	0.86		0.04	0.04	0.92	
				Bandwidth = 1000 Hz				
labial	0.78	0.11	0.12	0.75	0.97	0.01	0.01	0.82
alveolar	0.13	0.84	0.03		0.11	0.87	0.02	
velar	0.05	0.31	0.63		0.25	0.12	0.63	

Table 3.7: Confusion matrices for syllables containing [e] and [o].

text will cue bilabials. If it turns out that the transitions are perceived as flat because of the spectral reduction, this may further cue the perception of either alveolar or bilabial place of articulation.

However, it was shown that the burst peak frequency alone was less able to predict responses to new stimuli than a representation of the release burst that encoded global spectral shape. This may be because a single peak frequency is a relatively poor representation and perhaps other detailed spectral cues would offer better predictions of listeners' responses.

An alternative explanation for the pattern of confusions in listeners' responses might be that the compactness of the burst spectral peak is no longer perceptible in 1 kHz bandwidth noise-vocoded samples. This may also cause responses to shift to either [d] or [b] depending on the vowel context in which it occurs which may influence whether the burst has an upward or downward sloping spectral tilt.

In any event, it is possible that the perception of the release burst alone dominates the identification of stop place of articulation in noise-vocoded stimuli. The next chapter attempts to factor out the influence of the release burst by presenting listeners with noise-vocoded isolated bursts and formant transitions.

3.4 Discussion and conclusions

In Sec. 3.2, it was established that some detailed formant information is potentially recoverable from noise-vocoded speech with bandwidths up to 1000 Hz. However, although detailed spectral measurements performed much better than gross spectral shape properties for undistorted speech, it was shown that they were much less robust in the context of reduced frequency resolution when used in an automatic classification task.

Although it was shown that the usefulness of detailed spectral cues, such as burst peak and formant frequencies, cannot be ruled out in the perception of noise-vocoded speech, it was found in a perceptual experiment that gross spectral shape features in the form of Mel-frequency cepstral coefficients provided better predictions of listeners' responses. This was the case for both the unbiased formant tracking procedure as well as for the method of formant extraction which biased the performance of the model towards greater robustness in the context of reduced frequency resolution.

It was also found that at least one detailed spectral feature, notably the burst peak frequency, is potentially very well preserved at even the widest filter bandwidths. Nevertheless, a gross spectral shape representation of the relase burst was better able to predict listeners' responses in the hybrid-cue model. However, it was even suggested in Sec. 3.3, that perception of the release burst alone may explain qualitative differences in confusion patterns for velars in different vowel contexts. It may be the case then, that a detailed spectral model of formant transitions is still adequate for *burstless* stimuli. This issue is addressed in the following chapter.

The next chapter gives a closer examination of the behavior of formant frequency measurements in noise-vocoded stimuli and attempts to use these improved measures to model listeners' data. In addition, the perception of isolated bursts and formant transitions is compared to evaluate the hypothesis that burst perception dominates listener responses to noise-vocoded full syllables.

Appendix: MATLAB program for noise vocoder

```
function y = nvocoder(x, m, noise)
%NVOCODER Noise-excited channel vocoder.
%
%   Y = NVOCODER(X, M, FLAG) returns an M-channel noise-vocoded
%   transformation of the input signal vector X. The number of
%   channels in the vocoder includes both positive and negative
%   frequencies - i.e., a signal with sampling frequency of
%   16 kHz processed by a 32 channel noise-vocoder will have
%   individual channel bandwidths of 500 Hz each.  The input
%   variable FLAG indicates whether each channel is substituted
%   with signal-correlated noise. If FLAG is 0, spectral
%   information within individual subbands is preserved and the
%   resulting output signal Y is a reconstruction of the subband
%   decomposition.

xlen = length(x);
if mod(xlen,m), x(ceil(xlen/m)*m) = 0; end
chanlen = length(x)/m;

% replace full signal with signal correlated noise.
if m == 1
  y = x.*sign(rand(length(x), 1)-.5);
  return
end

beta = 5;
```

```
n = 134;
ord = n*m;
nwin = n+1;
hwin = n/2;

% h = low-pass filter
h = zeros(1,nwin*m);
h(1:ord+1) = m*fir1(ord, 1/m, kaiser(ord+1, beta));

sampidx = 1:chanlen;
chanidx = [hwin; (m-1:-1:1)'*(chanlen+nwin)+1];
y = zeros(chanlen+nwin, m);
y(sampidx(ones(m,1),:)+chanidx(:,ones(chanlen,1))) = x;

% polyphase decomposition of y
x = y;
for i = 2:m, x(:,i) = fftfilt(h(i:m:end), y(:,i)); end

chanlen = chanlen+1;
npos = fix(m/2)+1;

% gate out first samples to preserve overall length of output.
x = x(hwin+1:end-hwin,:);
f = fft(x, [], 2);

% signal-correlated noise
if noise
  f(:,1:npos) = f(:,1:npos).*sign(rand(chanlen,npos)-.5);
  f(:,npos+1:end) = conj(f(:,npos-~mod(m,2):-1:2));
end

% inverse of polyphase/fft decomposition
x = zeros(chanlen+nwin-1, m);
x(1:chanlen,:) = real(ifft(f, [], 2));

y = zeros(size(x));
y(hwin+1:chanlen+hwin,1) = x(1:chanlen,1);
for i = 2:m
  y(:,i) = fftfilt(fliplr(h(i:m:end)), x(:,i));
end

x = y(hwin+1:end-hwin,:);
y = reshape(flipud(x'), m*(chanlen), 1);
y = y(m:end-m+mod(xlen-1,m));
```

Chapter 4

The perception of noise-vocoded prevocalic stop bursts and formant transitions

4.1 Introduction

In Chap. 3, it was shown that, although some formant frequency information is *potentially* available to listeners in noise-vocoded speech up to 1 kHz bandwidth, these parameters were not able to predict subjects' identifications as well as gross-spectral-shape features in the form of Mel-frequency cepstral coefficients. In addition, it was shown that MFCCs were much more robust than detailed spectral features in the automatic classification of spectrally reduced stimuli.

There were, however, two factors that may have played a role in the poorer performance of detailed spectral measures in both the modeling of listeners' responses and the automatic classification experiment. Firstly, both of the methods used to extract formant frequencies from noise-vocoded speech were probably less than adequate. For example, the formant tracker implemented in the speech analysis program PRAAT makes certain assumptions that may not be true for anything but undistorted speech stimuli—*i.e.*, that formants are continuous, have relatively narrow bandwidths, and are relatively close to the expected average values observed for naturally produced tokens. None of these properties can necessarily be attributed to noise-vocoded speech. While it could be argued that this is evidence against the perception of formants in such spectrally degraded stimuli, it should be pointed out that none of these *ad hoc* strategies for formant tracking are intended to represent a model of perception, and if a detailed-spectral-cue theory is to be believed, an alternative approach to formant extraction in noise-vocoded speech must be taken.

The second method of formant frequency estimation selected formants from candidate LPC poles on the basis of their proximity to frequency estimates obtained from the original, unprocessed signals. While this biased the behavior of the model to greater robustness from the point of view of correct place-of-articulation classification, it may not have been able to capture the more systematic errors made by human listeners. It may be the case that the signal processing remaps formant frequencies in a completely nonlinear, but perfectly regular manner, thereby introducing a bias in the perception and categorization of spectrally reduced stimuli. This method of formant estimation is also a model based on extrinsic information—*i.e.*, the original formant frequencies are not available in the noise-vocoded tokens. Nevertheless, this approach was taken simply in order to give the benefit of the doubt to the perception of formant frequencies in spectrally distorted speech.

Because of the problems associated with the estimation of formant frequencies from noise-vocoded stimuli, the comparison of gross and detailed spectral measures may have favored global spectral shape parameters simply because of the ease with which MFCCs can be extracted from naturally produced speech. This is not to say, however, that listeners do not perceive place of articulation primarily on the basis of formant transition information—it simply means that the extraction of formants is relatively difficult and that the complex *ad hoc* procedures employed to do so can be viewed as somewhat controversial even for *clean-speech* stimuli.

Although these points do not explain the relatively poor performance observed for detailed spectral measures in the automatic classification task (in which percent correct classification was actually *overestimated* for the formant frequencies), the robustness of a simple pattern classification algorithm does not necessarily reflect the abilities of human listeners. In addition, a second factor may have worked against the detailed cues.

It was noted in the previous chapter that the representation of the release burst as a single peak frequency may not have been entirely adequate either. Several authors have suggested that bursts are considerably more complex, consisting of several prominent peaks (*e.g.*, Fischer-Jørgensen, 1954; Zue, 1976; Jongman and Miller, 1991). It was also shown that the pattern of listeners' confusions might be explained on the basis of the relative perceptual strength of the release burst alone: however, hypotheses derived from both detailed and gross-cue theories were given. It is nevertheless possible that listeners' responses were largely dominated by the perception of the burst. Therefore, because of the potentially inadequate representation given the release burst in terms of detailed spectral cues, it was felt that the perceptual influence of the consonant release should be treated separately from that of the formant transitions.

From a theoretical standpoint, treating the perception of detailed cues in the burst and formant transitions as distinct issues may be more appropriate, since competing theories of burst perception are not nearly as controversial as those that address the perception of the formant transitions themselves. For example, while many authors have proposed that the multiple spectral prominences found in release bursts can be generalized by a few simple patterns based on broadly defined templates (*e.g.*, Stevens and Blumstein, 1978; Blumstein and Stevens, 1979, 1980; Lahiri, Gewirth, and Blumstein, 1984), there are virtually no vocal proponents of a perceptual model in which the release burst is represented by simple detailed spectral features, such as the frequency of the most prominent spectral peak. While Winitz, Scheib, and Reeds (1972) attempted to correlate burst-peak center frequencies with listeners' identifications of gated voiceless releases, it appears that they included the aspiration portion in these measurements. Therefore, the perceptual results were very likely correlated with voiceless formant transitions. Although Smits, ten Bosch, and Collier (1996b) also modeled the perception of prevocalic stop consonants using detailed spectral cues, they found that the overall spectral tilt and the amplitude of the mid-frequency peak performed slightly better in predicting listeners' responses to isolated bursts. This latter observation was supported by results obtained in the previous chapter, where it was shown that a hybrid model, which used gross spectral shape features for the burst and formant frequencies for the vocalic portion, was also slightly better able to predict listeners' responses than the strict, detailed-cue model in which the burst was represented by its peak frequency. Other studies have presented empirical evidence in support of the role of a single spectral peak in the perception of stop bursts. However, they all used synthetic speech stimuli, in which the release bore little resemblance to naturally produced tokens (*e.g.*, Liberman, Delattre, and Cooper, 1952; Ainsworth, 1968). Because so few detailed cue theories for the perception of consonant burts have been proposed, it is perhaps better to focus attention on the formant transitions as a possible source of perceptually relevant detailed spectral information, if detailed cues are to be given a reasonable chance of success.

Most proponents of detailed spectral features as cues to place of articulation identification refer mainly to formant transitions as potential perceptual correlates (*e.g.*, Liberman *et al.*, 1967; Nearey and Shammass, 1987; Lindblom, 1990; Sussman *et al.*, 1998). The controversy between spectral shape *vs.* formant frequency also extends into the vowel perception literature. For example, it has been suggested that formants that are relatively close in frequency may not be perceived as distinct spectral features (*e.g.*, Chistovich and Chernova, 1986) and spectral-shape models of vowel perception that incorporate this effect have also been proposed (Hermansky, 1990). This can be contrasted with analyses that treat formant frequencies as independent acoustic properties (*e.g.*, Nearey and Assmann, 1986; Hillenbrand and Nearey, 1999). It was decided, therefore, to focus attention on the perception of the voiced formant transitions of prevocalic stop consonants alone, as the comparison between detailed and gross-spectral-shape features in this segment of speech has implications for many fields of speech perception research.

In summary, two potential problems were found in the experiments presented in Chap. 3. Firstly, the algorithms used for formant extraction may not have accurately represented listeners' perception of detailed spectral cues in noise-vocoded stimuli, and secondly, the representation of the release burst as a single peak frequency may have been relatively impoverished as an acoustic correlate to place-of-articulation perception. In this chapter, these two issues are addressed directly. In response to the first, a large continuum of synthetically produced speech stimuli, with known formant-frequency values, is noise vocoded and then analyzed to evaluate the patterns of distortion caused to formant frequencies. It is shown that this distortion is actually much more deterministic than had been assumed in the previous chapter. With a better understanding of the exact nature of this distortion, it is then possible to make predictions regarding listeners' categorizations of stop-consonant place of articulation in noise-vocoded speech, based on a detailed-cue model of formant perception.

In response to the second issue, noise-vocoded, *isolated* bursts and formant transitions are presented to listeners separately for identification of stop-consonant place of articulation. It is shown that the perception of formant-transitions alone is considerably less robust than that of gated bursts in spectrally reduced speech. This suggests that many of the responses obtained from subjects in the previous chapter may have been largely influenced by the perception of the release burst. Because it is believed that the burst was poorly modeled, detailed spectral measures may not have been given a fair chance of success against gross spectral shape parameters. Listeners' responses to isolated formant transitions were then modeled to evaluate the predictive ability of competing detailed and gross-cue models.

This chapter represents an attempt to give formant transitions as many advantages as possible in competing with spectral shape parameters. Because of the many complex factors that influence the extraction of these detailed spectral cues, it is felt that this may offset many of the natural disadvantages imposed on formant frequency measurements, such as the algorithm used to track them. However, if it is found that, under ideal conditions (which may or may not be realizable in actual speech perception), spectral shape parameters in the form of Mel-frequency cepstral coefficients are still better able to predict listeners' responses to noise-vocoded stimuli, then this result would represent a defeat of detailed spectral features as perceptual correlates to place of articulation in such spectrally degraded stimuli. Conversely, if the heavily optimized (but potentially unrealistic) formant frequency measures prove better able to predict subjects' responses, then we cannot rule out formants as potential cues.

Section 4.2 describes the experiment using noise-vocoded, synthetic stimuli. In Sec. 4.3, predictions from the synthesis experiment are exploited in an automatic classification task. Section 4.4 then describes the perception experiment using gated bursts and isolated formant transitions. This is followed by Sec. 4.5, which briefly discusses any conclusions that can be drawn from these experiments.

4.2 Synthesis experiment

As mentioned previously, the specific nature of the distortion of detailed spectral cues caused by noise vocoding is not known. However, a more detailed understanding of the effects of such spectral distortion may lead to better predictions of subjects' responses to noise-vocoded stimuli based on a detailed-cue model of perception. For example, it may be the case that the vocoding of vocalic formants remaps their frequency values in a perfectly predictable manner which is, in turn, reflected by systematic errors in listeners' perception of such spectrally distorted stimuli. An analysis of this distortion may then lead to a more optimal model of subjects' response errors based purely on detailed spectral cues in the form of vocalic formant transitions.

With the goal of obtaining a better representation of formants in spectrally reduced speech, it was decided to synthesize a large number of four-formant stimuli which were then processed by a noise vocoder. Measured formant frequencies from spectrally processed stimuli were then compared to the parameters used to generate the original tokens. Such a simulation is able to produce an estimate of the overall variability of the effects of spectral distortion as well as the magnitude and direction of the distortion itself. If it is found that the distortion of formant frequencies is relatively consistent in that the observed *variability* of the noise-vocoded formant measurements is relatively small compared to the *magnitude* of the distortion, then this simulation can provide better insight into the nature of the feature transformation involved. If this is the case, a remapping of the stimulus properties between unprocessed and noise-vocoded stimuli will have been obtained, which can then be used to predict listeners' errors in the perception of such spectrally reduced speech.

Section 4.2.1 describes the synthetic stimuli used in this simulation, while Sec. 4.2.2 defines the procedure used to measure formants from their noise-vocoded counterparts. Section 4.2.3 gives the results of this simulation.

4.2.1 Stimuli

Stimuli were generated using a MATLAB implementation of the cascade branch of the KLATT80 speech synthesizer (Klatt, 1980). A 10 second train of impulses was generated having $f_0 = 130$ Hz at a sampling rate of 8 kHz. This impulse train was filtered by a single pole centered at 0 Hz with a −3 dB bandwidth of approximately 77 Hz and a single zero at 1500 Hz with a −3 dB bandwidth of 6 kHz. This conforms to the synthesis of the default glottal source as implemented by the FORTRAN code in Klatt (1980).

Stimuli were synthesized with four formants having fixed bandwidths of 50 Hz, 70 Hz, 110 Hz, and 240 Hz for F_1–F_4 respectively. The third and fourth formant frequencies were also held fixed at 2665 Hz and 3300 Hz. 870 such stimuli were synthesized with F_1 frequencies ranging from 250 Hz to 810 Hz in 29 steps of 20 Hz each, and F_2 ranging from 920 Hz to 2370 Hz in 30 steps of 50 Hz each. This can be compared with the absolute ranges of F_1 and F_2 obtained from male speakers as described in Sec. 3.2.3, which were found to be 258–812 Hz and 916–2366 Hz respectively, including both the vocalic onset and vowel measurements. The mean value of F_3 from these stimuli was also 2665 Hz. Although F_2 and F_3 could have been varied parametrically in a similar manner, it was felt that variation in any two formants would illustrate the potential interactions that may occur. In addition, there was considerably more overlap in the ranges of F_2 and F_3 for male speakers (916–2366 Hz and 2056–3435 Hz for F_2 and F_3 respectively), while the smallest empirical differences between F_2 and F_3 formant frequencies were actually much larger (332 Hz) than those found in these synthetic stimuli (110 Hz).

Synthetic speech was thus produced by filtering the glottal source with the formants. For each stimulus synthesized, noise-vocoded tokens were generated at 250 Hz, 500 Hz, 1000 Hz, and 2000 Hz bandwidths. In total, there were 4 × 870 = 3480 10 second, noise-vocoded stimuli.

4.2.2 Formant measurements

Each 10 second synthetic stimulus was pre-emphasized by a factor of 0.96 and was then segmented into 800 25 ms overlapping frames with an overlap rate of 12.5 ms. A Kaiser window ($\beta = 5$) was applied to each section. A 9th order autoregressive model (LPC) was then estimated for each of the 800 frames for each stimulus using the modified covariance method (Kay, 1988) and the first three complex pole frequencies were estimated. However, if fewer than three such roots were found in the autoregressive series, the available candidates were assigned to the lowest formants in order, and the remaining formant frequencies were interpolated linearly from adjacent frames.

In order to adjust for minor errors in formant assignment, frequencies estimated for either F_1 or F_3 were permitted to be reassigned to F_2 if this was found to decrease the absolute difference between F_2 frequencies in the two adjacent, overlapping segments. However, only formant frequencies in nonadjacent frames were allowed to be reassigned in such a manner; therefore, gross errors in formant assignment were retained. Nevertheless, the total number of such reassignments was very small: 0, 6, 2789, and 24 049 segments out of a total of 696 000 such segments in each of the 250 Hz, 500 Hz, 1000 Hz, and 2000 Hz bandwidth conditions respectively. Thus, only 3.5% of the total number of estimated F_2 frequencies had to be adjusted in the 2 kHz bandwidth condition. This procedure is therefore similar to that proposed by Markel and Gray (1976).

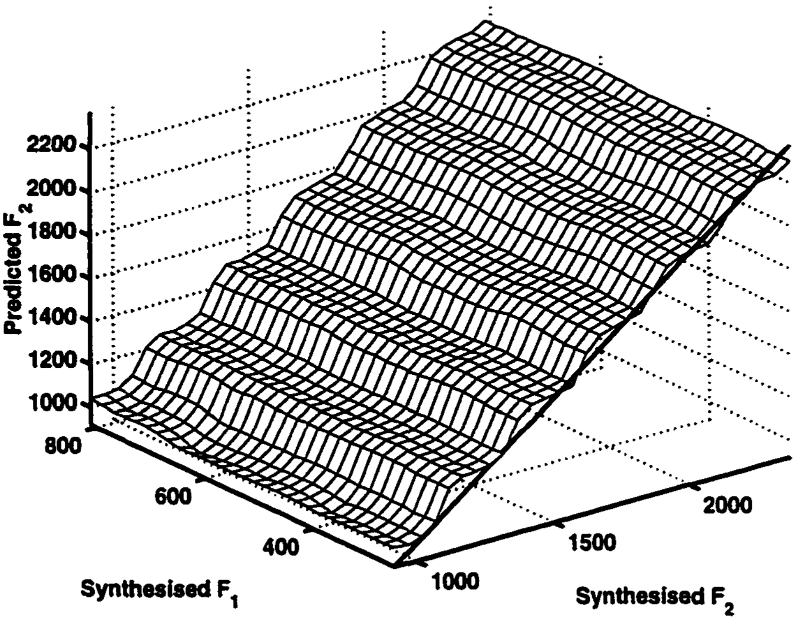


Figure 4.1: Measured F_2 formant frequencies for the 250 Hz bandwidth noise-vocoded synthetic stimuli. The thick line in front indicates the expected F_2 without spectral distortion.

4.2.3 Results

Figures 4.1, 4.2, 4.3, and 4.4 show the estimated F_2 formant frequencies from the noise-vocoded stimulus set as a function of the F_1 and F_2 synthesis parameters for the 250 Hz, 500 Hz, 1000 Hz, and 2000 Hz bandwidth conditions respectively. The dark diagonal lines also indicate the actual F_2 synthesis parameter. Three things can be observed from these surface plots: firstly, the estimate of the synthetic F_2 parameter from noise-vocoded stimuli is quite good at the narrower bandwidths 250 Hz, and 500 Hz. Secondly, there does not appear to be a large effect for F_1 in any of the plots. In addition, the surfaces have a very smooth and regular shape—particularly for the narrower bandwidths.

Although the relative smoothness, that can be seen particularly in Figs. 4.1 and 4.2 for the 250 Hz and 500 Hz bandwidth conditions, would seem to indicate that the variability found between individual measurements was relatively small, this is not necessarily the case; the standard error of the means becomes smaller and smaller with increased replications. The variance between measurements was therefore evaluated directly. The root-mean-squared errors from an analysis of variance based on the observations was calculated assuming either a cell-means model, an F_1 and F_2 main effects model, or a basic F_2-means model. These values are presented in Table 4.1 on p. 95. Although F-tests were performed using the mean-squared errors, the degrees of freedom in the denominator (695 130) were so large that it virtually guaranteed significance for all comparisons. This does not, however, give an impression of the magnitudes of these effects.

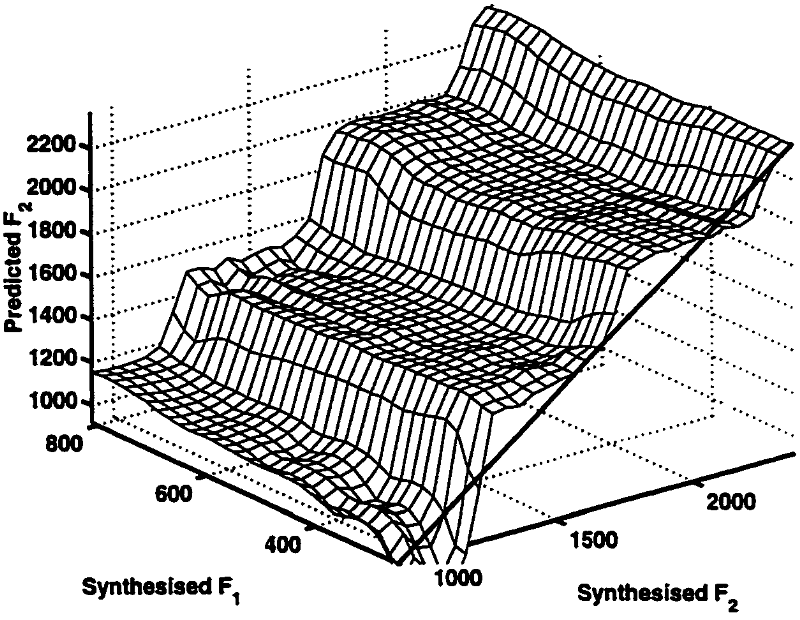


Figure 4.2: Noise vocoded F_2's at 500 Hz bandwidth

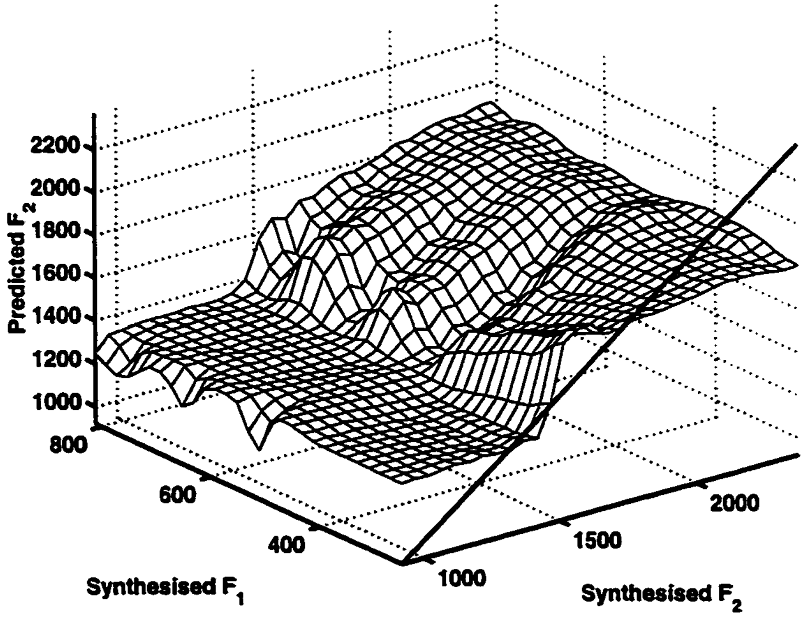


Figure 4.3: Noise-vocoded F_2's at 1000 Hz bandwidth

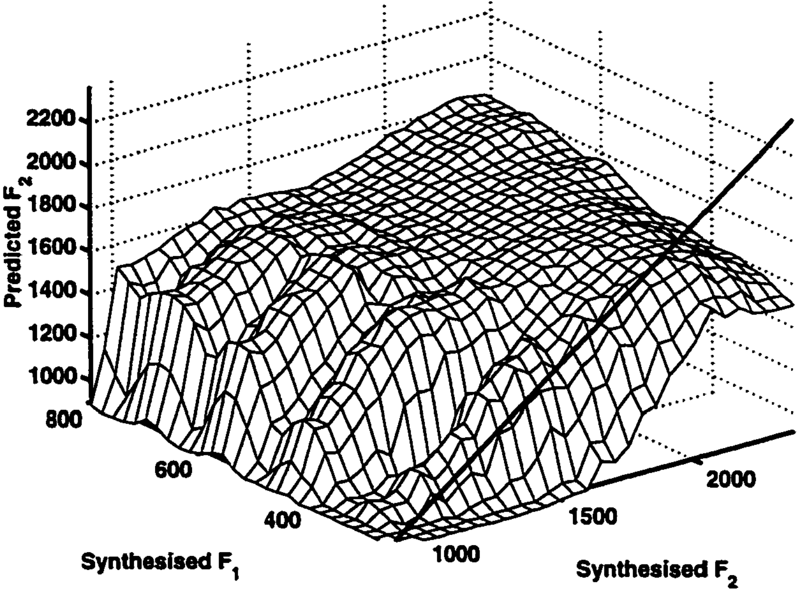


Figure 4.4: Noise-vocoded F_2's at 2000 Hz bandwidth

The largest model in Table 4.1 is based on the experimental error and represents the rms value assuming a full cell-means model:

$$y_{ijk} = \bar{\alpha}_{i..} + \bar{\beta}_{.j.} + \bar{\gamma}_{ij.} + e_{ijk} \tag{4.1}$$

where $\bar{\alpha}_{i..}$ is the mean of the measured F_2 formant frequencies for each level of F_1 ($i = 1\ldots29$), $\bar{\beta}_{.j.}$ is the mean for each level of F_2 ($j = 1\ldots30$), $\bar{\gamma}_{ij.}$ is the mean for each level of F_1 and F_2, and e_{ijk} is the observation error. The root-mean-squared error for the full cell-means model is therefore:

$$\text{rms} = \left(\frac{1}{29\cdot30\cdot800}\sum_i\sum_j\sum_k e_{ijk}^2\right)^{\frac{1}{2}} \tag{4.2}$$

These values are relatively small for the narrower bandwidths. In addition, it can be seen that removing the interaction term F_1F_2 ($\bar{\gamma}_{ij.}$ in Eq. 4.1) and the main effect term F_1 ($\bar{\alpha}_{i..}$) from the models does not increase the rms error appreciably, suggesting that the surfaces in Figs. 4.1, 4.2, 4.3, and 4.4 are reasonably well represented by the means of F_2 across levels of F_1.

The root-mean-squared *difference* between the measured noise-vocoded F_2 formant frequency and the actual synthesis parameter was approximately 56 Hz, 142 Hz, 260 Hz, and 404 Hz for the 250 Hz, 500 Hz, 1000 Hz, and 2000 Hz bandwidth conditions respectively. We would expect these values to be relatively small if formant frequency information is well preserved in noise-vocoded speech; these estimates can be compared with the those obtained by the method of nearest-pair formant extraction used in Sec. 3.2. By comparison with Fig. 3.10 on p. 71, which gives the rms

↓parameters	rms (Hz)	rms (Hz)
bandwidth→	250 Hz	1000 Hz
$F_1 + F_2 + F_1F_2$	33.05	107.93
$F_1 + F_2$	33.65	116.72
F_2	33.89	120.19
null	407.12	245.63
bandwidth→	500 Hz	2000 Hz
$F_1 + F_2 + F_1F_2$	69.16	281.70
$F_1 + F_2$	84.51	297.05
F_2	85.46	328.86
null	393.96	398.02

Table 4.1: Root-mean-squared error for models of vocoded F_2. The values give the rms error for the model predicting the F_2 formant frequency values measured from noise-vocoded stimuli.

difference between formant measurements from unprocessed stimuli and the nearest pole frequency from an LPC analysis of the corresponding noise-vocoded segments, rms values found here for the 250 Hz and 500 Hz bandwidth conditions are very similar. However, in the 1000 Hz and 2000 Hz bandwidth conditions, the rms estimates obtained by picking the nearest pole are much smaller. This suggests that the modeling procedure which biased noise-vocoded formant frequencies towards higher correct classification rates by virtue of their similarity to unprocessed measurements probably did not accurately represent these spectral properties in noise-vocoded speech. This may explain the poor goodness-of-fit statistics estimated from the modeling of listeners' responses. For example, although the rms difference between F_2 frequency measurements and synthesis parameters was approximately 260 Hz for 1000 Hz bandwidth noise-vocoded stimuli, the standard deviation of the measurements themselves for each level of F_2 was only 120 Hz (Table 4.1). This indicates that much of the distortion caused by noise vocoding can be attributed to a shift or remapping of the frequency values and that the method of nearest-pair formant matching was not a good representation of noise-vocoded formants.

Figure 4.5 on the following page also illustrates the mapping of F_2 synthesis parameters onto the noise-vocoded formant space. This figure shows side views of the surface plots presented in Figs. 4.2 and 4.3. The left-hand plots, which illustrate the main effects for F_1 and F_2 synthesis parameters on the measurements of 500 Hz noise-vocoded stimuli, also indicate that the influence of F_1 is relatively small. The noise-vocoded F_2 generally falls into one of three narrow bands of frequencies spaced approximately 500 Hz apart. The center frequencies of the 500 Hz bandwidth channels are, in fact, 500 Hz, 1000 Hz, *etc.*, suggesting that the poles of noise-vocoded stimuli are drawn towards the centers of individual subbands.

The side views of the F_2 map for the 1000 Hz bandwidth condition, on the right-hand side of Fig. 4.5 on the next page, is considerably more complex and hints at a larger effect for the F_1 synthesis parameter. The center frequencies of the channels in the 1 kHz vocoding condition are 1000 Hz, 2000 Hz, *etc.*, so there is no particular explanation for why these measurement values fall into two relatively narrow bands at

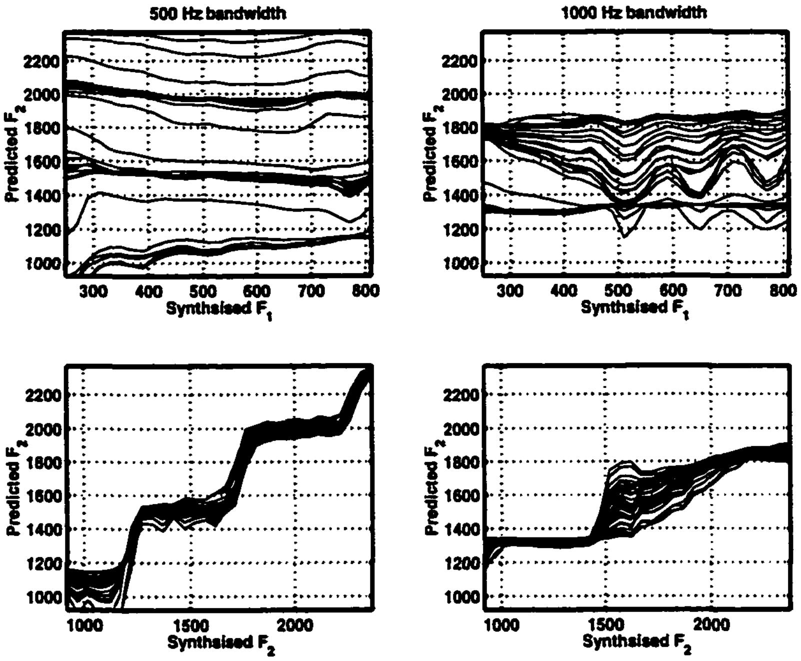


Figure 4.5: Side views of Figs. 4.2 on p. 93 and 4.3 on p. 93. In the top graphs, individual lines represent levels of F_2, while in the bottom graphs, lines represent levels of F_1.

approximately 1300 Hz and 1800 Hz. However, a shift does occur at 1500 Hz which is the frequency boundary *between* two subbands where the majority of F_2 measurements change from 1300 Hz to 1600–1800 Hz. This again suggests that formant estimates are perhaps biased towards the subband center frequencies.

Figure 4.6 on the following page shows the power spectral density of a synthetic vowel [æ] (bottom-most solid line) generated with formant frequencies of 620 Hz, 1660 Hz, and 2430 Hz for F_1, F_2, and F_3 respectively (remaining synthesis parameters were the same as in Sec. 4.2.1). The power spectral densities of the same stimulus, noise vocoded at both 500 Hz and 1000 Hz bandwidths, are also given as solid lines. As expected, the power spectra of the noise-vocoded signals appear as a discrete series of plateaus which represent the energy within dynamic range of each of the channels. In addition, the estimated LPC spectra for the original stimulus and the 500 Hz and 1000 Hz bandwidth noise-vocoded tokens are presented in the figure as dashed lines.

At 500 Hz vocoder bandwidth, it is clear that the first three formant peaks are well represented in the noise-vocoded power spectrum, which show local dominance of energy in the three subbands centered at 500 Hz, 1500 Hz, and 2500 Hz. The tendency appears to be for LPC poles to be drawn towards the centers of these high-energy bands as was illustrated in Fig. 4.5. However, at a vocoder bandwidth of 1000 Hz, the limited spectral resolution is unable to completely represent three distinct peaks. A band of energy between 500 Hz and 1500 Hz draws the LPC estimate of F_2 downward in frequency. Because the channel containing the frequency of the actual synthetic F_2

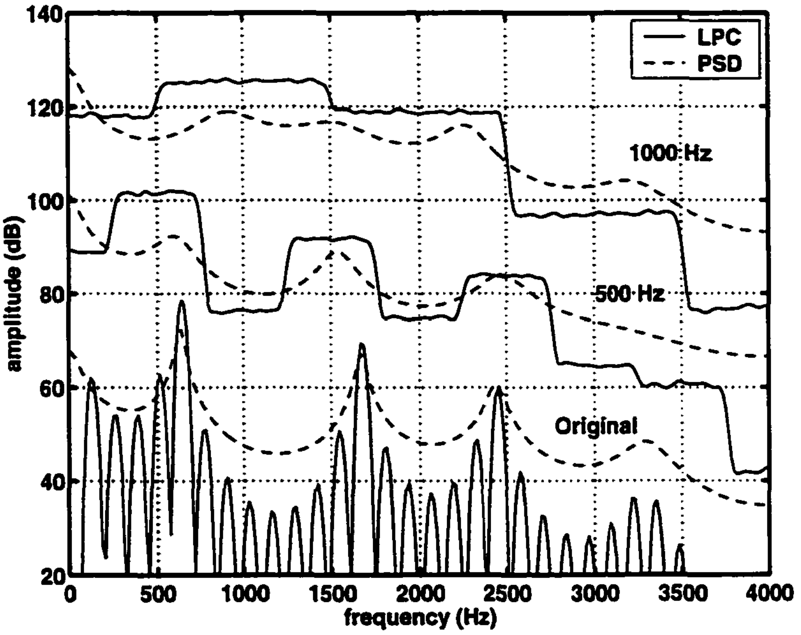


Figure 4.6: Power and LPC spectra of [æ] for the original synthetic stimulus and the corresponding stimuli noise-vocoded at 500 Hz and 1000 Hz.

also contains F_3, the estimated second formant is pushed downwards toward the higher energy subband for the 1000 Hz bandwidth noise-vocoded stimulus.

What happens to the LPC analysis of noise-vocoded speech when the frequencies of two formants are within 500 Hz of each other? We might expect the representation to break down. Figure 4.7 on the following page shows power spectral density functions of similar stimuli in which the F_1 and F_2 synthesis parameters were much closer in frequency (790 Hz and 970 Hz respectively) which results in a vowel that sounds roughly like [ɑ]. Only two subbands show local spectral dominance in the power spectral density of the 500 Hz bandwidth noise-vocoded stimulus. In addition, two of the LPC poles fall within the dynamic spectral range of a single channel. The result of this is that these two formant frequencies have not diverged greatly from their original synthesis values, illustrated by the peaks in the bottom-most dashed line. Because the channel spanning the range of frequencies 750–1250 Hz is considerably higher in amplitude, two poles are used to model the single high-energy plateau in order to match the relative spectral balance between channels.

This is perhaps not an unreasonable account of what might happen in perception according to a detailed-cue theory. A distinct F_1 and F_2 is neither plainly visible in the *original* synthesized stimulus illustrated in Fig. 4.7, and yet it sounds very much like [ɑ] and not some other vowel that might be heard if the peak at approximately 750 Hz were instead perceived as a *single* formant. It could be argued that this peak is, in fact, perceived as two distinct formants because of the high amplitude and wide bandwidth of the single prominent spectral peak, even though they are spectrally merged. Thus, the same reasoning that would lead to the assignment of F_1 and F_2 to the single

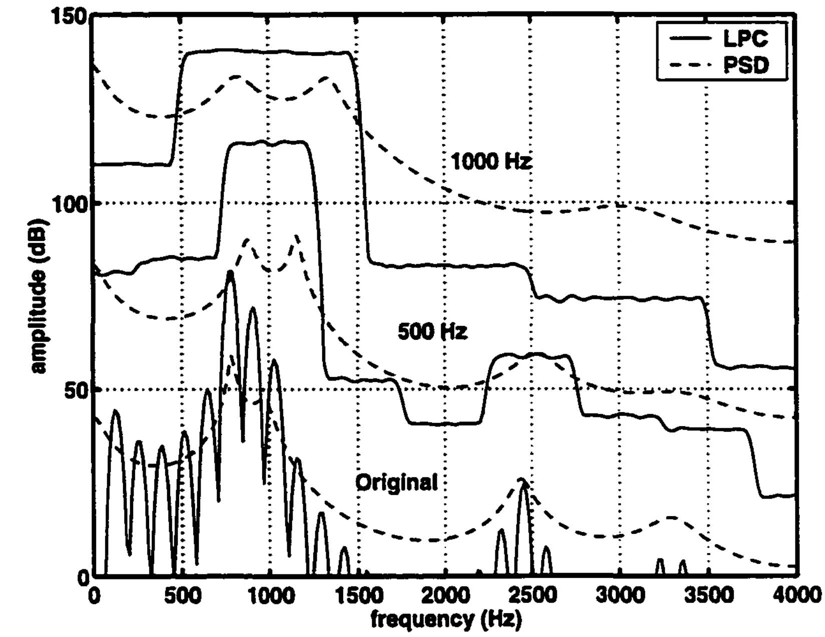


Figure 4.7: Power and LPC spectra of synthetic [a]

spectral prominence in the undistorted stimulus can also then be applied to the noise-vocoded signal as well. There are, however, differing opinions on how very closely spaced formants might be perceived in the perception of vowels. The alternate views are, nevertheless, roughly analogous to the opposition between gross and detailed-cue theories of stop consonant perception (see Sec. 4.1).

The merging of two formants within a single channel also occurs in the 1000 Hz bandwidth vocoded stimulus. It is interesting to note that the estimated value of the noise-vocoded F_2 in this example is 1298 Hz which was identified as one of the quantal regions in the distorted F_2 space for the 1000 Hz bandwidth condition (Fig. 4.5 on p. 96). The pole is unable to occupy the center frequency of the channel because it is competing with the pole representing F_1.

The next section evaluates the amount of stop-consonant place-of-articulation information that is potentially available in noise-vocoded formant transitions given the observed consistency of formant measurements from spectrally distorted, synthetic speech.

4.3 Automatic classification experiment

In the previous section, it was shown that the magnitude of the distortion of F_2 from the unprocessed to noise-vocoded synthetic stimuli was relatively large in comparison to the variability of the measurements themselves. This indicates that the method of nearest-pair formant extraction used in the modeling of listeners data in Chap. 3 may have biased the detailed-cue perceptual model toward lower goodness-of-fit scores, as it was probably not able to capture some of these systematic effects of noise-vocoding on formant frequency values. This was further illustrated by the fact that the rms difference between the measured F_2 formant frequencies from the distorted stimuli and the original F_2 synthesis parameters was much larger than that predicted by the nearest-pair algorithm for vocoding bandwidths of 1000 Hz and 2000 Hz. This suggests that formant frequency values in such spectrally degraded signals are consistently biased in some way. For example, in many cases, it was illustrated that the poles from the LPC analysis were biased towards the centers of individual subbands.

Nevertheless, the nearest-pair formant frequency estimation procedure actually biased the results of automatic classification experiments towards higher correct identification rates (Sec. 3.2.6). It would therefore be interesting to assess how the measurement of formant frequencies from distorted *synthetic* speech would fair in a similar test. This would again provide an upper bound on the potential availability of place-of-articulation information in noise-vocoded stimuli as well as the perceptual robustness of detailed spectral cues. Therefore, an automatic classification experiment was designed which exploited the hypothesized regularity of the mapping from undistorted to noise-vocoded formant frequencies.

4.3.1 Stimulus parameters

The stimuli used in this classification task are the same as those used in Sec. 3.2.1: 3 consonants × 9 vowels × 12 speakers = 324 stimuli. The signals were then noise-vocoded at 250 Hz, 500 Hz, 1000 Hz, and 2000 Hz.

Between one and ten Mel-frequency cepstral coefficients were measured from 25 ms segments centered at the onset of the vocalic formant transitions as well as the point 60 ms following. Each segment was weighted with a 25 ms Hamming function.

For the unprocessed stimuli, formant frequencies were also estimated at the onset of voiced formants and the point 60 ms following, in accordance with the procedure described in Sec. 3.2.3. The procedure for estimating formant frequencies from the noise-vocoded stimuli was considerably more complicated. In analogy with the automatic classification experiment presented in Sec. 3.2.6, formant frequency information from the original unprocessed stimuli was used to estimate parameters from the spectrally reduced tokens.

For each unprocessed stimulus, the median glottal period, as estimated from the glottal epoch detection procedure described in Sec. 3.2.3, was used to generate a 25 ms train of impulses. This train of impulses was then filtered to have a −6 dB/octave downward slope in the manner described in Sec. 4.2.1. This synthetic glottal source was filtered by a bank of cascade resonators whose frequency and bandwidth parameters were those estimated from the unprocessed stimuli. A 25 ms segment of synthetic speech corresponding to the formant onset and the onset + 60 ms point of each syllable was generated in such a manner, having the same frequencies and bandwidth parameters for the first four formants as those estimated from the original stimuli. These synthetic tokens were then noise-vocoded at 250 Hz, 500 Hz, 1000 Hz, and 2000 Hz, and the formant frequencies from the resulting stimuli were then measured using a 9th order LPC analysis via the modified-covariance method.

This procedure is expected to be an unbiased test of the maximum *potential* place-of-articulation information available in spectrally distorted speech signals. This experiment is therefore very similar to the one presented in Sec. 3.2.6. However, the problem of formant tracking is now circumvented by ensuring that the original synthetic stimuli

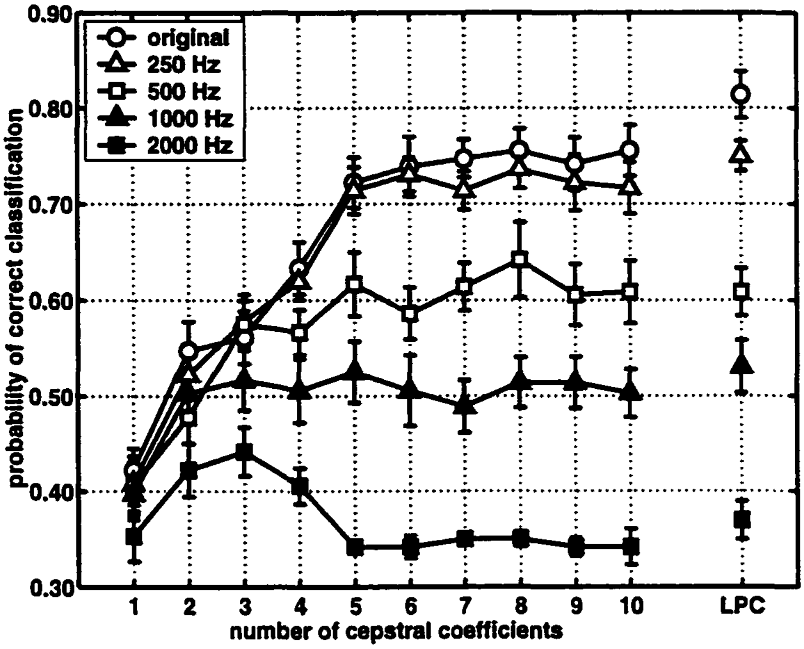


Figure 4.8: Probability of correct classification for MFCCs and LPC. See also Fig. 3.13 on p. 74.

contain exactly four formants so that the order of the LPC analysis could be minimized to estimate only four poles[1]. Thus, explicit formant tracking was not necessary.

The leave-one-speaker-out cross-validated percent correct classification rates were determined via a quadratic discriminant analysis (QDA) in a procedure identical to the one described in Sec. 3.2.6. The parameters of a QDA were estimated from the undistorted tokens produced by 11 speakers and these parameters were used to estimate the place of articulation of the tokens produced by the remaining speaker for both unprocessed and noise-vocoded stimuli. This procedure was repeated for all 12 speakers.

The spectral properties of the burst were ignored in this experiment. Therefore, the classification test is intended to reflect the identification of *burstless* formant transitions.

4.3.2 Results

The results of the automatic classification experiment are presented in Fig. 4.8. The correct classification rates for all models are, in general lower here than for the experiment presented in Sec. 3.2.6 (*cf.*, Fig. 3.13 on p. 74). The main reason for this decreased performance is the fact that the the spectral properties of the bursts were ignored in this test. This results in much lower correct classification rates for all the MFCC and LPC models.

[1] Occasionally, the LPC analysis results in fewer poles. If this was the case, the order of the analysis was increased by two until four poles were obtained.

Notably, however, we start to see significant differences in the spectral shape models between 250 Hz and 500 Hz vocoding conditions with more than three MFCCs. This is unlike the results from the classification experiment in which the spectral shape parameters for the release burst were also included; in that analysis, it was found that the 250 Hz and 500 Hz vocoding conditions showed nonsignificant differences with up to nine coefficients. Nevertheless, even without the burst, the correct classification results for the 250 Hz vocoding condition is largely indistinguishable from those for the original unprocessed stimuli.

It is interesting to note that the correct classification rates of undistorted stimuli obtained for the LPC models is significantly better than that for MFCCs. In addition, there appears to be no significant difference between LPC and MFCC models for the noise-vocoded stimuli. This suggests that a detailed-spectral model of classification is optimal for the categorization of stop place of articulation from isolated formant transitions. However, the results for the detailed spectral cues are biased in the sense that the speech stimuli used in this analysis were optimized for the measurement of exactly four formants and therefore required no complex formant tracking procedure. In contrast, stimuli used in the measurement of spectral shape parameters were not optimized in their favor, as MFCCs are already very easy to obtain. Although Mel-frequency cepstral coefficients could also have been measured from resynthesized speech, this presumably would have biased the comparison of results even further in favor of formant frequency information. For example, because the synthetic stimuli contained no energy above the fourth formant, this may bias the global spectral tilt, which is thought to be a very important acoustic cue at the onset of voiced formants (Lahiri *et al.*, 1984).

Otherwise, the relative comparisons between the spectral shape and detailed parameters are similar to those observed in Sec. 4.3.2: large differences are found between all processing conditions for the LPC measurements, while no differences in correct classification are found for MFCCs between unprocessed and distorted stimuli generated at the narrowest vocoding bandwidth. Based on gross-spectral-shape theory of stop consonant perception, we might expect human listeners to show no differences in correct classification performance between unprocessed and 250 Hz noise-vocoded burstless syllables, while for a detailed-cue theory, we might expect significant differences at even the lowest levels of distortion.

4.4 Perception experiment

The primary purpose of this experiment was to determine the relative perceptual robustness of isolated stop bursts and formant transitions. It was suggested in the previous chapter that much of listeners' ability to correctly identify the place of articulation of noise-vocoded stop consonants may be due to the robustness of release-burst spectral cues in the context of this type of reduced frequency selectivity. It may therefore be the case that the perception of isolated stop bursts by human listeners is far more robust than the perception of isolated formant transitions in noise-vocoded stimuli. This would support a theory which suggests that listeners attend primarily to detailed spectral cues, since the burst and vocalic portions are typically not treated differentially by gross-spectral-shape theories of speech perception (*e.g.*, Blumstein and Stevens, 1979).

It was suggested in the previous section that models based on spectral shape parameters in the form of Mel-Frequency cepstral coefficients are relatively insensitive to small amounts of spectral distortion, as it was shown that there was no significant differences in the rates of correct classification between unprocessed and 250 Hz noise-vocoded stimuli. However, if subjects show a significant decrease in identification rates for even the smallest levels of spectral distortion, this would lend even further evidence in support of a detailed-spectral theory of stop-consonant perception.

4.4.1 Subjects

Twelve native speakers of Western Canadian English were paid as subjects in this experiment. All subjects were either graduate or undergraduate students of Linguistics and none reported any hearing impairment.

4.4.2 Stimuli

Stimuli were the same as those used in Sec. 3.3.1 except that only syllables with an initial [b], [d], or [g] were used. In total, 3 consonants × 4 vowels × 12 speakers = 144 syllables were present in the initial stimulus set. However, in addition to whole syllables, release bursts and formant transitions were isolated and presented to listeners for identification as well. The location of the release burst for each stimulus was determined interactively using a waveform and spectrogram display. Once segmented from the full syllable, the tail-end of the release burst was tapered by the second half of a 10 ms Hamming window function. Similarly, the initial 5 ms of the isolated vocalic segment was tapered as well.

Full syllables, isolated bursts, and isolated formant transitions were noise-vocoded at four bandwidths: 250 Hz, 500 Hz, 1000 Hz, and 2000 Hz. Noise-vocoded and original, unprocessed stimuli were presented to listeners for identification of the prevocalic stop.

4.4.3 Procedure

Subjects were presented with unsegmented syllables, gated bursts, and isolated formant transitions in three separate sessions. Unprocessed and noise-vocoded stimuli were fully randomized within each session.

Unlike, the experiment described in Sec. 3.3.3, subjects were asked to choose from only [b], [d], and [g]. Each subject heard each stimulus only once. There were therefore 144 syllables × 5 processing conditions = 720 stimuli presented in each session.

Stimuli were resampled at 44.1 kHz and played through a Gina AD/DA at 16 bit quantization from a PC. Subjects listened to stimuli in a sound-attenuated room at a comfortable listening level.

4.4.4 Results

Figures 4.9, 4.10, and 4.11 show the probabilities of correct stop identification as a function of signal processing condition, for each subject in the experiment, for whole syllables, isolated bursts and vocalic segments respectively. The thick dashed line at 33% indicates chance performance, while the dark error bars indicate ± one standard error for the subject means. Narrow dashed lines indicate individual subject performance.

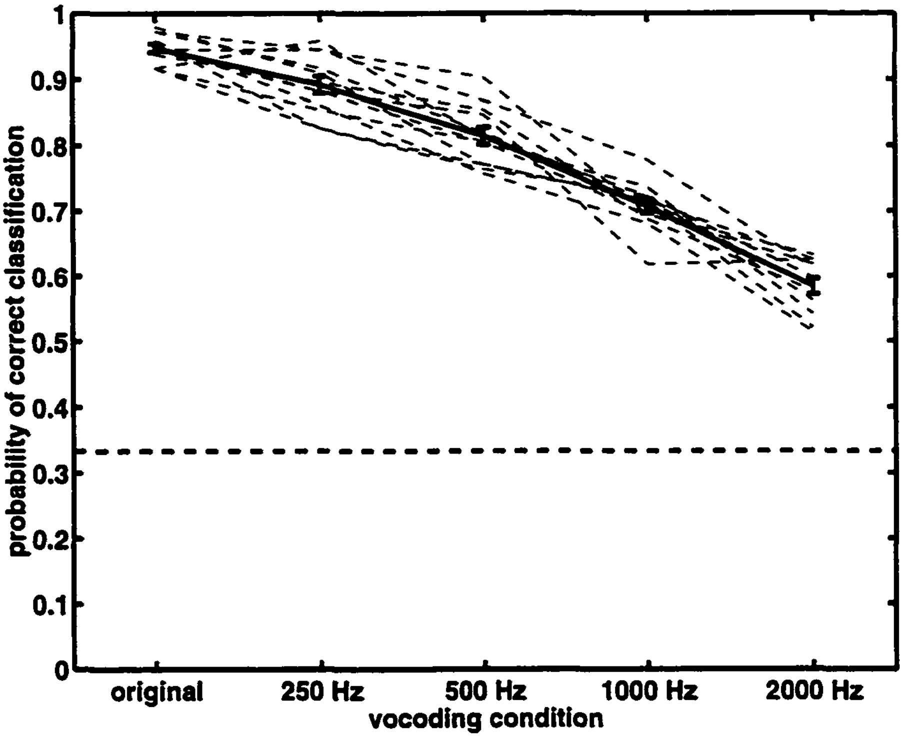


Figure 4.9: Probability of correct stop identification: whole syllables. The error bars indicate ± one standard error for the means across subjects, while the thick dashed line indicates chance performance. The light dashed lines indicate individual subject scores.

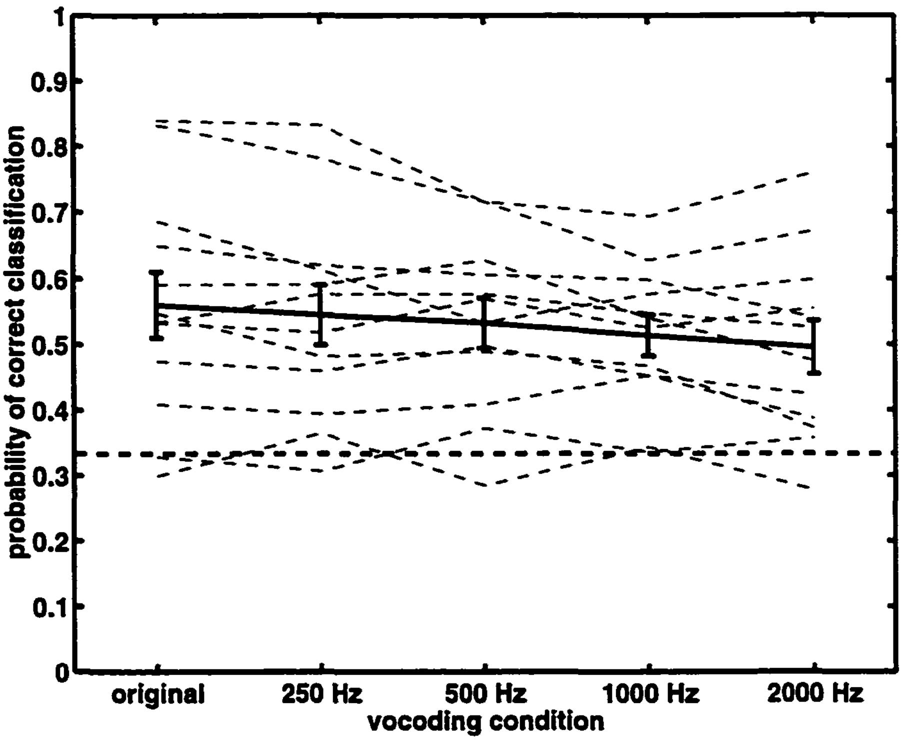


Figure 4.10: Probability of correct stop identification: isolated bursts

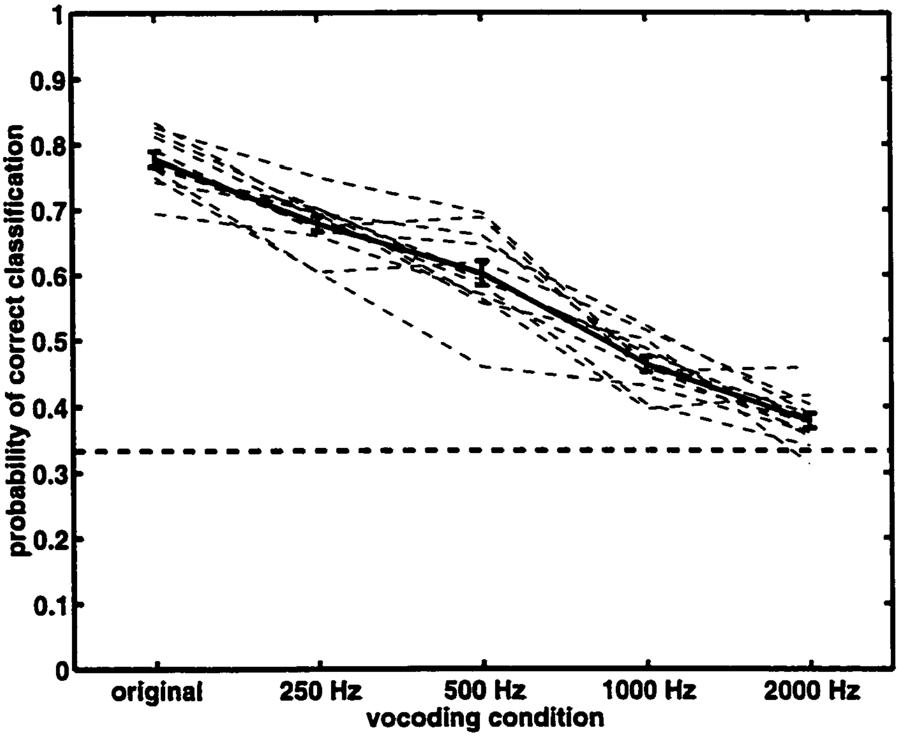


Figure 4.11: Probability of correct stop identification: isolated vocalic segments

Figure 4.10 on p. 104 indicates that, although there was a wide range in performance between individual subjects, overall means did not appear to change significantly across the signal processing conditions for the perception of isolated bursts. Nevertheless, the correct recognition rates for the unprocessed stimuli are considerably lower than that reported in similar experiments of isolated burst perception. For example, Smits *et al.* (1996a) obtained an average recognition rate of 73.6% for short-lag VOT Dutch stops.

One possible explanation for these lower identification rates may have to do with the number of speakers used in the stimulus set. For example, Smits *et al.* (1996a) blocked stimuli by speaker while here, all twelve speakers were completely randomized within each session. It has been shown that the correct identification of words in noise is affected by the speaker blocking condition, suggesting that the additional cognitive load associated with speaker normalization may drain resources needed for the identification task itself (Mullennix *et al.*, 1989). Similarly, Malécot (1958), Kewley-Port (1982), Schouten and Pols (1983), and Krull (1990) used stimuli produced by only one speaker in experiments of isolated burst perception. Although Bonneau *et al.* (1996) used three speakers in a completely randomized design, they were all male. There have apparently been no experiments addressing the perception of gated, voiced bursts using speakers of both genders, completely randomized within sessions as is the case here.

However, the fact that correct identification performance does not change significantly between processing conditions indicates that isolated burst perception is remarkably robust despite the reduced frequency resolution.

Figure 4.11 on the preceding page shows that, although the perception of isolated formant transitions does not reach chance performance, even at the 2000 Hz bandwidth condition, there is almost a linear relationship between the log of the vocoding bandwidth and the percent correct identification. Extrapolation of this loglinear trend suggests that completely random identification performance might be achieved at vocoding bandwidths just slightly wider than 2 kHz.

It would appear from these figures that, although correct stop place identification from whole syllables (Fig. 4.9 on p. 103) remains significantly above chance levels—even at the widest vocoding bandwidths—much of this perceptual robustness is probably related to the correct identification of the release burst itself. In contrast, the perception of isolated vocalic segments (Fig. 4.11 on the preceding page) decreases rapidly across the five processing conditions. However, the categorization of burstless stimuli is still significantly above chance even at 2000 Hz vocoding bandwidth ($t = 4.14; \alpha < .001$, 11 degrees of freedom), yet performance is significantly worse for the isolated vocalic segments between the unprocessed and 250 Hz signal processing condition ($t = 7.88$; $\alpha < .001$ based on paired difference t-test with 11 degrees of freedom). This actually supports the view that listeners attend primarily to detailed spectral cues, as it was shown in the previous section that spectral shape cues appear to be largely insensitive to this level of distortion.

Tables 4.2, 4.3, and 4.4 give the percent correct identification for each subject and for each processing condition. Between each of the columns indicating identification scores are the significance levels of the McNemar tests (Fleiss, 1981), in which n_A is compared to the binomial distribution $B(n_A + n_B, \frac{1}{2})$ where n_A and n_B are the number of errors made in condition A not made in condition B and *vice versa* (Ripley, 1996). However, in order to evaluate the significance of the differences in processing conditions *across* subjects, a t-test with 11 degrees of freedom is performed on the the modified McNemar statistics $\pm\sqrt{m}$ (Eq. 2.2 on p. 43). These show the same results as indicated

	unproc.		250 Hz		500 Hz		1000 Hz		2000 Hz
S	pcc	α	pcc	α	pcc	α	pcc	α	pcc
A	0.94	0.62	0.94	0.01*	0.87	0.02*	0.78	0.00**	0.61
B	0.92	0.97	0.96	0.00**	0.81	0.04*	0.72	0.01*	0.60
C	0.98	0.00**	0.91	0.00**	0.77	0.15	0.72	0.00**	0.52
D	0.92	0.01*	0.83	0.11	0.77	0.14	0.72	0.01*	0.58
E	0.95	0.01*	0.88	0.03*	0.81	0.00**	0.69	0.06	0.62
F	0.94	0.02*	0.88	0.19	0.85	0.00**	0.62	0.60	0.63
G	0.96	0.00**	0.83	0.05	0.76	0.20	0.72	0.00**	0.56
H	0.97	0.11	0.94	0.12	0.90	0.00**	0.69	0.12	0.63
I	0.92	0.05	0.85	0.14	0.81	0.10	0.74	0.00**	0.54
J	0.96	0.00**	0.86	0.01*	0.76	0.07	0.68	0.00**	0.51
K	0.97	0.02*	0.92	0.00**	0.82	0.01*	0.72	0.04*	0.63
L	0.94	0.13	0.90	0.15	0.85	0.00**	0.69	0.01*	0.58

Table 4.2: Correct identification scores for full syllables. The values between the columns give the probability of significant differences between adjacent processing conditions—*i.e.*, n_A is compared to $B(n_A + n_B, \frac{1}{2})$, where n_A and n_B are the number of errors made in condition A not made in condition B and *vice versa*.
$^*\alpha < 0.05$; $^{**}\alpha < 0.01$

by the errorbars in the plots: highly significant differences between processing conditions for full syllables ($t = 4.65$; $t = 6.85$; $t = 5.68$; $t = 7.45$; $\alpha < .001$ for all tests), and for isolated formant transitions ($t = 8.32$; $t = 4.62$; $t = 7.44$; $t = 5.33$; $\alpha < .001$ for all tests). However, for gated bursts, no significant difference was found between adjacent processing conditions ($t = 1.26$; $t = 1.06$; $t = 1.42$; $t = 1.06$; *n.s.*). In fact, based on the McNemar tests shown in Table 4.3 on the next page, there was only one significant difference found for a single listener in one signal processing condition for the perception of isolated bursts.

4.4.5 Modeling of listeners' responses

It was suggested in Sec. 4.2 and Sec. 4.3 that voiced formant transitions may contain a substantial amount of place-of-articulation information, even in very distorted stimuli. In Sec. 4.2 it was established that noise-vocoded formant frequencies formed regular patterns as a function of the synthesis parameters used to gererate the corresponding stimuli. In Sec. 4.3, it was shown that this pattern—which can be thought of as a formant-frequency mapping function from an unprocessed to noise-vocoded feature space—could be used to successfully classify place of articulation in such distorted stimuli. This was done be resynthesizing tokens based on empirical measurements from unprocessed tokens, noise-vocoding the synthetic stimuli and then taking formant frequency measurements from the distorted signals. The problem is, however, that the robust information purportedly contained within such noise-vocoded stimuli is still essentially hidden in distorted, naturally produced stimuli—*i.e.*, the automatic classification experiment was forced to rely on information *extrinsic* to the stimuli actually being classified.

It has already been established in Chap. 3 that a formant tracking algorithm de-

	unproc.		250 Hz		500 Hz		1000 Hz		2000 Hz
S	pcc	α	pcc	α	pcc	α	pcc	α	pcc
A	0.30	0.92	0.36	0.11	0.28	0.89	0.34	0.13	0.28
B	0.53	0.82	0.58	0.56	0.58	0.35	0.55	0.40	0.53
C	0.65	0.33	0.62	0.44	0.61	0.50	0.60	0.19	0.54
D	0.53	0.43	0.52	0.88	0.57	0.21	0.53	0.80	0.55
E	0.69	0.09	0.61	0.08	0.53	0.83	0.58	0.70	0.60
F	0.55	0.13	0.48	0.61	0.49	0.40	0.47	0.07	0.37
G	0.59	0.56	0.59	0.79	0.63	0.06	0.54	0.16	0.47
H	0.47	0.45	0.46	0.76	0.50	0.25	0.45	0.35	0.42
I	0.41	0.45	0.39	0.64	0.41	0.80	0.45	0.16	0.39
J	0.33	0.40	0.31	0.92	0.37	0.30	0.34	0.70	0.36
K	0.83	0.14	0.78	0.07	0.72	0.05	0.63	0.83	0.67
L	0.84	0.50	0.83	0.00**	0.72	0.34	0.69	0.94	0.76

Table 4.3: Correct identification scores for gated burst.$^{*}\alpha < 0.05$; $^{**}\alpha < 0.01$

	unproc.		250 Hz		500 Hz		1000 Hz		2000 Hz
S	pcc	α	pcc	α	pcc	α	pcc	α	pcc
A	0.77	0.02*	0.67	0.67	0.69	0.00**	0.44	0.13	0.38
B	0.74	0.16	0.70	0.00**	0.56	0.00**	0.40	0.70	0.42
C	0.81	0.00**	0.69	0.03*	0.58	0.00**	0.40	0.12	0.34
D	0.76	0.00**	0.60	0.69	0.62	0.02*	0.51	0.02*	0.40
E	0.82	0.00**	0.70	0.02*	0.59	0.04*	0.49	0.02*	0.35
F	0.76	0.07	0.69	0.00**	0.56	0.17	0.50	0.00**	0.31
G	0.79	0.00**	0.67	0.37	0.65	0.01*	0.52	0.01*	0.39
H	0.76	0.07	0.70	0.01*	0.60	0.01*	0.48	0.06	0.38
I	0.69	0.23	0.66	0.04*	0.57	0.01*	0.45	0.59	0.46
J	0.75	0.00**	0.60	0.01*	0.46	0.35	0.43	0.11	0.36
K	0.83	0.00**	0.69	0.28	0.66	0.00**	0.47	0.10	0.39
L	0.83	0.04*	0.75	0.15	0.69	0.00**	0.46	0.03*	0.36

Table 4.4: Correct identification scores for isolated formant transitions.
$^{*}\alpha < 0.05$; $^{**}\alpha < 0.01$

signed for clean speech performs relatively badly when used to predict listeners' responses to noise-vocoded stimuli. However, it was felt that a number of factors may have weakened its potential. Firstly, the burst peak frequency was included in the analysis. It was subsequently shown that a hybrid model, which included spectral shape parameters for the burst, improved its predictive performance substantially. However, it was also shown that a completely *ad hoc* method of formant extraction based on the proximity of candidate LPC poles to formant measurements from the original unprocessed counterparts also performed much more successfully. The problem with this latter model is that, it too relies on an extrinsic source of information and, therefore, cannot represent a model of actual perception. It was included in the analysis because it was felt that it was perhaps a better test of a model's *potential* ability to predict listeners' responses to such stimuli given ideal conditions, whether such conditions were physically realizable or not.

The problems associated with the spectral representation of the burst are not addressed in the modeling of these data: only burstless, isolated formant transitions are considered. While this may avoid the first issue, how can we capture the regularities observed for noise-vocoded formants in Sec. 4.2? When analyzing the noise-vocoded synthetic stimuli, the main source of extrinsic information exploited by the formant extraction procedure was the number of formants used to generate the unprocessed synthetic tokens. Often, formant estimation algorithms model extra poles which are then discarded via the tracking algorithm (*e.g.*, Talkin, 1987). However, only a very simple formant tracking algorithm was needed for the distorted, synthetic speech stimuli in Sec. 4.2, and no formant tracking algorithm was used at all in the automatic classification experiment in Sec. 4.3. Because there were no extraneous poles returned by the LPC analysis, the process of assigning candidate poles to formants was straightforward—*i.e.*, formant tracking was not necessary.

Therefore, in modeling subjects' responses to burstless stop consonants, the procedure used to extract formants from noise-vocoded stimuli will assume that listeners are somehow aware of the number of formants that are expected in the noise-vocoded stimuli. Therefore, the problem of formant tracking in noise-vocoded stimuli is circumvented by estimating the minimal expected number of poles in an LPC analysis. Again, this provides an upper bound on the potential ability of listeners to perceive formants in noise-vocoded speech.

All tokens were resampled at 8 kHz. Formant frequencies were then estimated for two 25 ms segments centered at either the onset of voiced formants or the point 60 ms following this. Each segment was pre-emphasized by a factor of 0.96 and weighted by a 25 ms Hamming function. Linear predictor coefficients were then estimated via the modified coviariance method (Kay, 1988). However, the order of the analysis was based on the number of formants that were present in the original, unprocessed stimulus—*i.e.*, two times the number of formants plus one to adjust for the overall spectral tilt. The roots of the autoregressive polynomial were then solved and the three complex poles with the smallest angles were then assigned in order to the first three formants.

Two hundred cross-validation trials were performed in the manner described in Sec. 2.3.5 and as illustrated in Fig. 2.7 on p. 47—*i.e.*, six subjects were drawn, without replacement, from the total set of responses and a loglinear model was estimated based on their responses to stimuli produced by six speakers, also randomly drawn without replacement. This model was then used to estimate *a posteriori* probabilities of place of articulation for stimuli produced by the remaining six speakers. The percent modal agreement (Eq. 2.8 on p. 48) and root-mean-squared error (Eq. 2.9) were evaluated on the basis of responses given by the remaining six subjects.

At each iteration, separate models were estimated for both the formant parameters and cepstral coefficients. For the formant model, both the squares and cross-products of the formants estimated from the vocalic onset and vowel segments were used. A total of 15 terms were thus used in the model estimation: F_{1_o}, F_{2_o}, F_{3_o}, F_{1_v}, F_{2_v}, F_{3_v}; the squares of all these parameters; and the cross-products $F_{1_o}F_{1_v}$, $F_{2_o}F_{2_v}$, and $F_{3_o}F_{3_v}$. These additional squares and cross-products were included in order to capture the co-variation observed in Figs. 3.11 and 3.12. It is therefore assumed that the covariance between corresponding onset and vowel formant frequencies is independent for each place of articulation. First formant frequency parameters were also included because the omission of burst features

Several MFCC models were considered, in which the number of coefficients ranged from one to six. Each analysis window was 25 ms in length with the first centered at the onset of the vocalic formant transitions after the preceding burst had been removed. The second window was centered 60 ms following this. Each frame was tapered by a 25 ms Hamming window and the Mel-frequency cepstral coefficients for each segment were estimated.

At each iteration, the cross-validated percent modal agreement and root-mean-squared error between listeners' responses and the formant model was compared with that for the MFCC models. If the spectral-shape model is significantly closer to the true model underlying listeners' responses, then we expect that MFCC parameters will obtain a higher pma (and lower rms) in at least 95% of the trials. The converse would be true if formant frequency estimates are, in fact, closer—*i.e.*, MFCC parameters would obtain higher pma (and lower rms) in less than 5% of the trials.

Instead of pooling the data from all processing conditions, they were instead included in successive, nested datasets. The first model consisted of data from only the undistorted tokens, the second consisted of data from undistorted and 250 Hz bandwidth noise-vocoded stimuli, the third also included 500 Hz bandwidth vocoded stimuli, *etc.*

Table 4.5 on the following page gives the proportion of cross-validation trials in which the MFCC models attain better goodness of fit than the three-formant model as a function of both the number of coefficients and the number of processing conditions included in the analysis. As the table indicates, no significant differences were observed between the formant model and the spectral shape models for distorted stimuli with vocoding bandwidths of either 250 Hz or 500 Hz, except for MFCC models with only one coefficient in each frame. It would appear that the formant frequency model is probably at least as adequate as any of the MFCC models up to 500 Hz vocoding bandwidth. Although only nonsignificant differences were observed for the unprocessed and 250 Hz and 500 Hz noise-vocoded stimuli, the tendency appears to be that formants perform somewhat better than spectral shape parameters for the original, unprocessed stimuli, while MFCCs are slightly better able to fit listeners' responses to the distorted stimuli. In fact, in 500 Hz vocoding condition, the rms goodness-of-fit statistic for the MFCC model with five coefficients is better than that for the formant frequencies in 94% of the trials. While this is not a significant result, it should be noted that the number of estimable parameters in this spectral-shape model was only 22 *vs.* 32 for the formant frequency model. Thus the spectral-shape model was able to generalize to new stimuli and new responses in the cross-validation procedure with many fewer degrees of freedom. At 1000 Hz bandwidth and up, the spectral shape model performs

	unproc only		up to 250 Hz		up to 500 Hz		up to 1000 Hz		up to 2000 Hz	
#	pma	rms	pma	rms	pma	rms	pma	rms	pma	rms
1	0.01	0.06	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01
2	0.01	0.06	0.17	0.22	0.22	0.25	0.45	0.55	0.47	0.55
3	0.03	0.12	0.35	0.58	0.45	0.61	0.80	0.90	0.92	0.95
4	0.06	0.18	0.45	0.68	0.63	0.81	0.91	0.98	0.96	0.99
5	0.23	0.33	0.68	0.79	0.82	0.94	0.99	0.99	1.00	1.00
6	0.14	0.20	0.58	0.73	0.77	0.86	0.96	0.95	1.00	1.00

Table 4.5: Proportion of trials in which MFCCs obtained better goodness of fit.

significantly better than formant frequencies with as few as four cepstral coefficients.

Figure 4.12 on the next page gives the cross-validation percent modal agreement for each of the MFCC models. The figure seems to indicate that the spectral shape model is actually better able to predict listeners' responses to noise-vocoded stimuli than to the original unprocessed tokens. However, this difference is not significant: the pma score for the 250 Hz bandwidth stimuli exceeded that for the original, unprocessed tokens in only 65%, 83%, 69%, 77%, 67%, and 71% of the trials for between one and six coefficients respectively. There is, nevertheless, the tendency for the inclusion of noise-vocoded stimuli to improve the predictive power of the spectral shape model.

These pma scores can be compared with those obtained for the formant model, which were 0.647, 0.610, 0.595, 0.530, and 0.498 for the nested processing condition datasets from unprocessed through 2000 Hz bandwidth respectively. It should be noted that although the *differences* in goodness-of-fit statistics between different models is expected to be unbiased, the absolute values of these estimates are, in fact, biased downwards because only one quarter of the data is used to estimate the generalized linear model (Shao, 1993, see also Sec. 2.3.6).

The number of formants present in the unprocessed stimulus is still a source of information extrinsic to the noise-vocoded stimuli. What we have not considered here is whether listeners might be able to determine the number of potential formants in noise-vocoded speech. It is not known, therefore, how much this added source of perceptual variability might affect the estimates of formant frequencies in these stimuli. In this analysis, the assumption was explicitly made that this information is somehow able to be determined under some ideal (but, again, perhaps impossible) conditions. It is certainly possible, however, that this information could be available to listeners in the less distorted stimuli. For example, it was shown in sec 4.3 that the relative prominence of individual subbands was able to represent individual formants in 500 Hz bandwidth noise-vocoded stimuli and the number of formants in the dynamic range of the speech stimuli could perhaps be determined from these spectral prominences. It was even implied that very high amplitude subbands could be modeled as two merged formants, which could also assist listeners in perception under a detailed-cue hypothesis. However, at 1000 Hz bandwidth, the assumption that listeners may be able to determine the expected number of formants present in the stimulus seems unlikely. Nevertheless, it was found that spectral shape parameters were significantly better than the formant frequency model when 1000 Hz vocoded stimuli were included in the dataset.

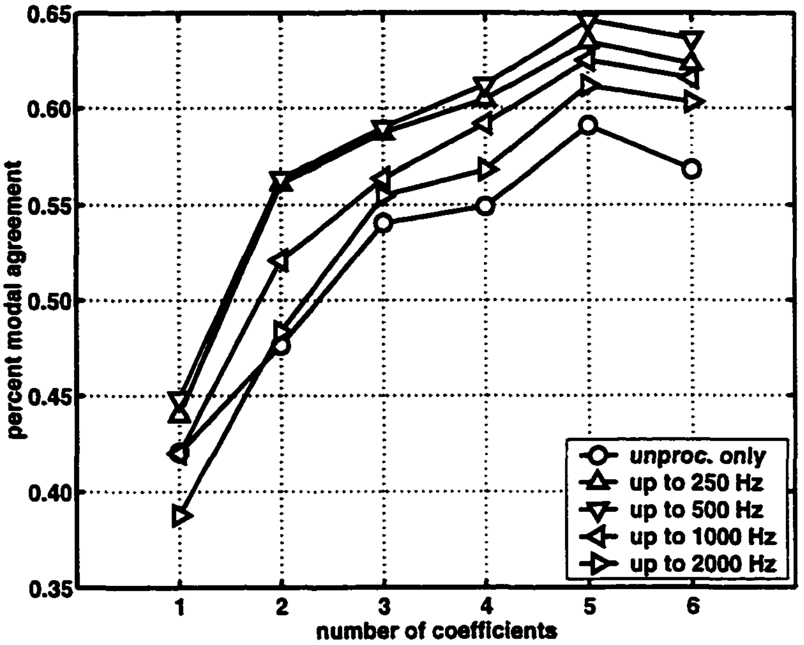


Figure 4.12: Cross-validation percent modal agreement for MFCC models

4.4.6 Conclusions

It would appear that, despite the best efforts to represent formants in noise-vocoded speech, spectral-shape parameters in the form of Mel-frequency cepstral coefficients are still better able to predict listeners' identifications of spectrally distorted stimuli with noise-vocoding bandwidths of 1000 Hz and broader. However, detailed spectral features are still more optimal for unprocessed formant transitions.

4.5 Summary and conclusions

In Sec. 4.2, it was shown that the effects of noise-vocoding on synthetic speech stimuli were entirely systematic in that the variance of the measurements *within* levels of F_1 and F_2 synthesis parameters was relatively small compared to the magnitude of the distortion itself. Some general observations regarding the transformation between synthesis parameters and estimated noise-vocoded formant frequencies were made and it was suggested that the perception of noise-vocoded speech on the basis of formant frequencies could not be ruled out. It was also felt that *ad hoc* strategies of formant estimation may have biased modeling results against detailed spectral cues because of their inability to capture some of the systematic effects observed in this experiment.

The fact that much formant frequency information is *potentially* available in noise-vocoded speech was illustrated in the automatic classification experiment in Sec. 4.3. In this experiment, it was found that the correct identification of the place of articulation of noise-vocoded, burstless synthetic stimuli was comparable to rates obtained for spectral shape parameters in the form of Mel-frequency cepstral coefficients. However, it was noted that these results may be biased in favor of the detailed spectral measures, as the stimuli that were classified by the formant frequency discriminant analysis were resynthesized based on the formant parameters measured from unprocessed tokens. Because they contained only the first four formants, the need for an explicit formant tracker was circumvented, which avoided the increased variability associated with this procedure.

In Sec. 4.4, it was found that the correct identification of gated bursts by human listeners did not change significantly between bandwidth conditions. It was concluded that the perception of the isolated releases was quite robust, which may explain much of listeners' ability to correctly identify the place of articulation of noise-vocoded syllables at wide vocoder bandwidths. Therefore, it can be claimed that much of the perceptual robustness observed in simulations of electric hearing, which resemble the noise-vocoded stimuli used in this experiment, can largely be attributed to the perception of the burst alone (*e.g.*, Shannon *et al.*, 1995; Dorman *et al.*, 1997).

It was also found that there was a significant decrease in correct responses to *burstless* stimuli that were noise-vocoded at a bandwidth of 250 Hz. This supports the view that listeners may be attending to some detailed spectral information, as this result was predicted for the perception of detailed cues based on the observations made in the automatic classification experiment (Sec. 4.3). Conversely, a spectral shape model would have predicted no significant differences in perception to the least distorted, noise-vocoded stimuli.

Listeners' responses to these stimuli were then modeled using either detailed spectral measures, in the form of formant frequencies, or gross-spectral features, in the form of Mel-frequency cepstral coefficients. Based on the regularities observed for the vocoding of synthetic speech stimuli, it was predicted that some formant frequency information may be available in even wide-bandwidth noise-vocoded stimuli. In order to give as much benefit of the doubt to formant frequency measures as possible, it was decided to fix the order of the autoregressive model based on the number of formants that had been observed in the original, unprocessed stimuli below 4 kHz. This was done under the assumption that listeners may be able to estimate the number of formants present within this dynamic range under some kind of ideal conditions.

However, despite this effort to bias the predictive power of the formant frequency model, it was not only found that the spectral shape model still performed better at the wider bandwidth conditions, but that no significant improvement for either model was found for the narrower bandwidth, noise-vocoded stimuli. There was, however, a significant advantage for the formant model for the original, unprocessed stimuli.

It is concluded therefore, that despite efforts to model the distortion of formant frequencies in noise-vocoded tokens, these estimates were not sufficient to account for listeners' perception of these stimuli. It is also concluded that spectral shape parameters in the form of MFCCs provide significantly better fits to subjects' identifications of the more spectrally distorted segments. Conversely, a formant model is significantly better for undistorted speech.

These observations are roughly analogous to the results obtained by Lindholm *et al.* (1988), who found that hearing-impaired listeners were more likely to respond to conflicting-cue stimuli on the basis of gross spectral shape than normal-hearing subjects. The circumstances under which these results were obtained are, nevertheless, quite different: hearing-impaired listeners have presumably adapted to their hearing loss, while in this experiment, noise-vocoded and unprocessed signals were completely randomized within sessions and presented to normal-hearing listeners who had very minimal exposure to such stimuli.

Likewise, Shannon *et al.* (1995) and Dorman *et al.* (1997) trained listeners before collecting data on very wide-bandwidth noise-vocoded speech, presumably because these authors felt that a long adaptation phase was necessary. However, listeners in this experiment were capable of classifying *burstless*, noise-vocoded stimuli at significantly above chance levels for even the most distorted tokens. Prior adaptation cannot not explain this behavior.

In conclusion, while the perception of formants in speech with very low levels of spectral distortion is indeed very possible, it is unlikely that a formant only theory of speech perception can explain subjects' responses to very wide bandwidth noise-vocoded stimuli.

Chapter 5

Concluding remarks

In this dissertation, competing theories of stop-consonant place-of-articulation perception were evaluated on the basis of their ability to predict listeners' identification of spectrally and temporally distorted stimuli. The signal processing methods used to modify naturally produced tokens were chosen on the basis of their expected ability to effectively remove one set of spectral or temporal features from the stimuli. If it was found that the distortion had no effect on listeners' ability to identify relevant phonetic properties, it could then be assumed that the set of acoustic features that were found to remain unaffected by the signal processing must contain all of the important phonetic information needed to perceive such stimuli.

In reality, however, such results are never obtained. If the set of relevant acoustic features were known exactly, then an appropriate signal processing algorithm could be designed which reduced speech signals to only the minimally required phonetic information without any detriment to listeners' ability to identify such stimuli. Obviously, if these features were known, the experiments would not have to be performed in the first place; a method of spectral or temporal processing that completely preserves all of the relevant phonetic detail in speech was therefore not obtainable. If it is found then that the distortion causes some decrease in listeners' ability to identify the phonetic properties of the resulting signals, then complex statistical methods must be employed to evaluate whether a specific set of acoustic features is better able to predict listeners' responses than another.

This dissertation treats human speech recognition as statistical pattern classification (Nearey, 1997). Statistical methods can therefore be used to evaluate competing theories of speech perception on the basis of their ability to predict listeners' identification of stimuli that have been manipulated to isolate specific acoustic properties proposed by such theories. However, these perceptual models can also be compared with the *automatic* classification of similar speech stimuli.

The approach to the study of stop-consonant place-of-articulation perception taken in this dissertation has been to study the effects of spectral and temporal distortion on the ability of automatic classifiers to discriminate such speech sounds. These results are then compared to those obtained from human perception experiments. Automatic classification is useful in evaluating the *potential* ability of human listeners to identify distorted speech sounds. However, such tests can really only give a general impression of what to expect from true perceptual experiments, as it cannot be guaranteed that an automatic classification algorithm will discriminate speech categories in exactly the same way that humans do. Because training sets used to estimate the parameters of such discriminant analysis are necessarily limited to the availability of such data, they cannot normally be expected to perform as well as human listeners who have had considerably more exposure to speech stimuli. Nevertheless, such experiments are a useful first step in predicting the performance of human listeners in a perceptual task. Actual subjects' responses to such stimuli are then modeled based on explicit theories of speech perception to evaluate their ability to predict these data.

In Chap. 2, it was found that the distortion of the waveform envelope had no effect on listeners' ability to identify the place of articulation of short, gated stop bursts when the total duration of the stimulus was less than the voice onset time. It was therefore suggested that a purely *static* representation of the release burst up to 20 ms in duration was sufficient to represent all the important phonetic detail contained in these short stimuli.

The signal processing method employed in this experiment modified the waveform envelope of the original stimuli while preserving the long-term static spectral characteristics. Nevertheless, it was felt that this method may not have been directly related to the cognitive representation of place of articulation and, therefore, listeners' responses were modeled using one of two sets of acoustic cues. Ultimately, it was found that a model which incorporated *dynamic* spectral features was no better able to predict listeners' responses to these stimuli than a purely *static* representation that imposed no representation on the temporal organization of acoustic features. However, when the gated burst included at least one glottal epoch, it was then found that a purely static spectral model was inadequate. This suggested that the temporal organization of the burst and vocalic portions must be held distinct in the perception of prevocalic stops. This conclusion was, therefore, contrary to the assumptions made by Stevens and Blumstein (1978); and Blumstein and Stevens (1979, 1980) who claimed that the burst and vocalic segments formed an integrated spectral property regardless of the relative onset time of voicing, or whether the burst was actually present or not. However, it was suggested that the model proposed by Lahiri, Gewirth, and Blumstein (1984) could actually accommodate these results. In their model, they proposed that the relative change in spectral tilt between the burst and the onset of voicing could be used to distinguish [b] from [d].

The focus of the next two chapters was on the nature of *spectral* instead of *temporal* acoustic properties. Two distinct sets of spectral features have been proposed as potential acoustic correlates to the perception of stop-consonants place of articulation. These sets differ primarily in terms of the spectral resolution required to represent the relevant spectral features. In the case of gross-spectral-cue properties, such as the gross spectral shape or compactness of the mid-frequency peak, it was suggested that only a very broad resolution was required to accurately represent most of the global properties of stop consonants. The alternative view suggested that the most important acoustic correlates to place-of-articulation identification in prevocalic stops was found in relatively detailed spectral features such as the frequencies of the release burst peak and vocalic formant transitions. In contrast with gross-spectral-shape cues, it was hypothesized that these features required a much narrower spectral resolution.

It was suggested that the reduction of the effective spectral resolution of these speech stimuli may help distinguish the predictions proposed by these competing models. Because gross-spectral-shape features are broadly defined, it was predicted that such acoustic cues would be relatively robust in the context of reduced frequency selectivity characteristic of noise-vocoding, which reduced the spectral resolution of the speech signals. Conversely, it was predicted that detailed spectral measures would not exhibit the same level of robustness under the same signal processing conditions. Based on previous experiments involving the perception of noise-vocoded stimuli, it was hypothesized that listeners may, in fact, attend primarily global-spectral shape features, as high levels of correct recognition accuracy had been reported in simulations of cochlear-implant processing (Shannon *et al.*, 1995; Dorman *et al.*, 1997). Nevertheless, there was also evidence that speech perception was noticeably degraded at even very low levels of distortion, indicating that listeners may instead attend primarily to detailed spectral cues (Celmer and Bienvenue, 1987; ter Keurs *et al.*, 1992; Boothroyd *et al.*, 1996).

In simulation experiments, it was found that detailed spectral features, in the form of burst peak and formant frequencies, were perhaps not as fragile in the context of spectral distortion as might have been expected. Nevertheless, automatic classification experiments indicated that the ability of detailed spectral measures to correctly identify the place of articulation of noise-vocoded stimuli was severely limited in comparison to gross spectral shape features, which were represented by Mel-frequency cepstral coefficients. Results from a perceptual experiment in which the identification of unprocessed, naturally produced stimuli was compared with that of spectrally distorted stimuli, which were noise-vocoded at 500 Hz and 1000 Hz bandwidths, showed that subjects were indeed very sensitive to such levels of distortion. This was the result predicted in the automatic classification experiment using detailed spectral measures. However, when listeners' responses were modeled using either explicit detailed or gross-spectral features, it was found that the spectral shape parameters were better able to predict the data.

However, it was felt that there were many factors that sided in favor of the gross spectral measures. Formant frequencies are, in fact, relatively difficult to obtain from completely undistorted stimuli; noise-vocoding made these estimates even more problematic. In contrast, Mel-frequency cepstral coefficients are relatively easy to measure. In addition, it was also felt that the detailed spectral representation of the release bursts was inadequate and that the investigation of gross *vs.* detailed spectral measures should instead focus on the formant transitions, since the opposition between competing theories is much more contentious for this segment of speech. Nevertheless, the problem of isolating formants from noise-vocoded speech made it almost impossible to evaluate their perceptual significance.

Because of the natural disadvantages imposed on formant frequencies, it was decided to give as much benefit of the doubt as possible to detailed spectral measures under the assumption that a significant result in favor of gross-spectral shape features would be a more powerful indicator of their usefulness as perceptual correlates to place of articulation. In the first experiment, the correct classification performance of formant transitions was biased upwards by ensuring that the measured formant frequencies were as similar as possible to those obtained from the original, unprocessed tokens.

However, while this presumably improved the correct-identification performance of the detailed spectral model in the automatic classification task, it may not have been able to represent more systematic differences that exist between unprocessed and noise-vocoded conditions. Ultimately, this procedure did not prove to be successful in comparison with Mel-frequency cepstral coefficients in the modeling of listeners' data to noise-vocoded stimuli.

A more detailed analysis of the effects of noise-vocoding on formant frequency values was then performed. It was found that the magnitude of the distortion was, in general, larger than the variability of the measures themselves, suggesting that there were systematic changes in the noise-vocoded formants that were not captured by the procedure that artificially inflated the correct-identification performance of such stimuli. It was hypothesized that, because this consistent bias in formant frequency measures may have been reflected in the error patterns from listeners' responses, the predictive power of the formant model may have been underestimated.

However, when these systematic differences were incorporated into the measurement of formant frequencies from burstless, noise-vocoded speech, it was found that gross-spectral-shape features were *still* able to predict listeners' responses significantly better for stimuli that were noise-vocoded at bandwidths of 1000 Hz or wider. In addition, only nonsignificant differences were observed for stimuli noise-vocoded at narrower bandwidths.

While detailed spectral measures were nevertheless significantly better at predicting listeners' responses to *unprocessed* stimuli, it is clear that they are inadequate for the perception of noise-vocoded speech. This result suggests one of two things: that *both* spectral shape and formant frequencies are relevant for the perception of prevocalic stops, or that there is otherwise a better model that has not been considered. The first possibility could not be tested because of the limited number of stimuli and responses present in the dataset. However, the possibility cannot be overlooked. With regards to the second possibility, further work is needed to uncover possible additional acoustic sources of place-of-articulation information.

Bibliography

Ainsworth, W. A. (1968). "Perception of stop consonants in synthetic CV syllables," Lang. Speech **11**, 139–155.

Akansu, A. N. and Haddad, R. A. (1992). *Multiresolution Signal Decomposition* (Academic Press, Boston).

Allen, J. B. (1994). "How do humans process and recognize speech?" IEEE Trans. Speech Audio Process. **2**, 567–577.

Allen, J. B. and Rabiner, L. R. (1977). "A unified approach to short-time Fourier analysis and synthesis," Proc. IEEE **65**, 1558–1564.

Andruski, J. E. and Nearey, T. M. (1992). "On the sufficiency of compound target specification of isoalated vowels and vowels in /bVb/ syllables," J. Acoust. Soc. Am. **91**, 390–410.

Atkinson, A. C. (1981). "Likelihood ratios, posterior odds and information criteria," J. Econometrics **16**, 15–20.

Barry, W. J. (1984). "Place-of-articulation information in the closure voicing of plosives," J. Acoust. Soc. Am. **76**, 1245–1247.

Benkí, J. (1998). "Evidence for phonological categories from speech perception," Doctoral thesis, University of Massechusetts Amherst.

Blumstein, S. E., Isaacs, E., and Mertus, J. (1982). "The role of gross spectral shape as a perceptual cue to place of articulation in initial stop consonants," J. Acoust. Soc. Am. **72**, 43–50.

Blumstein, S. E. and Stevens, K. N. (1979). "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," J. Acoust. Soc. Am. **66**, 1001–1017.

Blumstein, S. E. and Stevens, K. N. (1980). "Perceptual invariance and onset spectra for stop consonants in different vowel environments," J. Acoust. Soc. Am. **67**, 648–662.

Bonneau, A., Djezzar, L., and Laprie, Y. (1996). "Perception of the place of articulation of French stop bursts," J. Acoust. Soc. Am. **100**, 555–564.

Boothroyd, A., Mulhearn, B., Gong, J., and Ostroff, J. (1996). "Effects of spectral smearing on phoneme and word recognition," J. Acoust. Soc. Am. **100**, 1807–1818.

Celmer, R. D. and Bienvenue, G. R. (1987). "Critical bands in the perception of speech signals by normal and sensorineural hearing loss listeners," in *The Psychophysics of Speech Perception*, edited by M. E. H. Schouten (Nijhoff, Dordrecht, the Netherlands), pp. 473–480.

Chistovich, I. A. and Chernova, E. I. (1986). "Inentification of one- and two-formant steady-state vowels: A model and experiments," Speech Commun. 5, 3–16.

Clark, H. H. (1973). "The language-as-fixed-effect fallacy: A critique of languate statistics in psychological research," J. Verbal Learn. Verbal Behav. **12**, 335–359.

Cole, R. A. and Scott, B. (1974a). "The phantom in the phoneme: Invariant cues for stop consonants," Percept. Psychophys. **15**, 101–107.

Cole, R. A. and Scott, B. (1974b). "Toward a theory of speech perception," Psychol. Rev. **81**, 348–374.

Cooper, F. S. (1950). "Spectrum analysis," J. Acoust. Soc. Am. **22**, 761–762.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., and Gerstman, L. J. (1952). "Some experiments on the perception of synthetic speech sounds," J. Acoust. Soc. Am. **24**, 597–606.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application* (Cambridge University Press, Cambridge, United Kingdom).

de Cheveigné, A. and Kawahara, H. (1999). "Missing-data model of vowel identification," J. Acoust. Soc. Am. **105**, 3497–3508.

Delattre, P. (1969). "Coarticulation and the locus theory," Studia Linguistica **23**, 1–26.

Delattre, P. C., Liberman, A. M., and Cooper, F. S. (1955). "Acoustic loci and transitional cues for consonants," J. Acoust. Soc. Am. **27**, 769–773.

Dorman, M. F. and Loizou, P. C. (1996). "Relative spectral change and formant transitions as cues to labial and alveolar place of articulation," J. Acoust. Soc. Am. **100**, 3825–3830.

Dorman, M. F. and Loizou, P. C. (1998). "The identification of consonants and vowels by cochlear implant patients using a 6-channel continuous interleaved sampling processor and by normal-hearing subjects using simulations of processors with two to nine channels," Ear Hear. **19**, 162–166.

Dorman, M. F., Loizou, P. C., and Rainey, D. (1997). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**, 2403–2411.

Dorman, M. F., Studdert-Kennedy, M., and Raphael, L. J. (1977). "Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues," Percept. Psychophys. **22**, 109–122.

Efron, B. and Tibshirani, R. (1997). "Improvements on cross-validation: The .632+ bootstrap method," J. Am. Statist. Assoc. **92**, 548–560.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap* (Chapman & Hall, New York).

Eimas, P. D. (1963). "The relation between identification and discrimination along speech and non-speech continua," Lang. Speech **6**, 206–217.

Fant, G. (1970). *Acoustic Theory of Speech Production: With Calculations Based on X-Ray Studies of Russian Articulations* (Mouton, The Hague, the Netherlands), 2nd ed.

Fant, G. (1973). "Stops in CV-syllables," in *Speech Sounds and Features*, edited by G. Fant (MIT Press, Cambridge, MA), pp. 110–139.

Fischer-Jørgensen, E. (1954). "Acoustic analysis of stop consonants," Miscellanea Phonetica **2**, 42–59.

Fischer-Jørgensen, E. (1972). "Tape cutting experiments with Danish stop consonants in initial position," Ann. Rep., Inst. Phonetics, Univ. Copenhagen **6**, 104–168.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions* (Wiley, New York), 2nd ed.

Forrest, K., Weismer, G., Milenkovic, P., and Dougall, R. N. (1988). "Statistical analysis of word-initial voiceless obstruents: Preliminary data," J. Acoust. Soc. Am. **84**, 115–123.

Fowler, C. A. (1986). "An event approach to the study of speech perception from a direct-realist perspective," J. Phonetics **14**, 3–28.

Fowler, C. A. (1994). "Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation," Percept. Psychophys. **55**, 597–610.

Fruchter, D. and Sussman, H. M. (1997). "The percpetual relevance of locus equations," J. Acoust. Soc. Am. **102**, 2997–3008.

Fujimura, O., Macchi, M. J., and Streeter, L. A. (1978). "Perception of stop consonants with conflicting transitional cues: A cross-linguistic study," Lang. Speech **21**, 337–353.

Gumpertz, M. and Pantula, S. G. (1989). "A simple approach to inference in random coefficient models," Am. Statistician **43**, 203–210.

Halle, M., Hughes, G. W., and Radley, J.-P. A. (1957). "Acoustic properties of stop consonants," J. Acoust. Soc. Am. **29**, 107–116.

Harris, C. M. (1953). "A study of the building blocks of speech," J. Acoust. Soc. Am. **25**, 962–969.

Harris, K. S., Hoffman, H. S., Liberman, A. M., Delattre, P. C., and Cooper, F. S. (1958). "Effect of third-formant transitions on the perception of the voiced stop consonants," J. Acoust. Soc. Am. **30**, 122–126.

Hermansky, H. (1990). "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am. **87**, 1738–1752.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.

Hillenbrand, J. M. and Nearey, T. M. (1999). "Identification of resynthesized /hVd/ utterances: Effects of formant contour," J. Acoust. Soc. Am. **105**, 3509–3523.

Hoffman, H. S. (1958). "Study of some cues in the perception of the voiced stop consonants," J. Acoust. Soc. Am. **30**, 1035–1041.

Jakobson, R., Fant, G., and Halle, M. (1965). *Preliminaries to Speech Analysis: The Distinctinve Features and Their Correlates* (MIT Press, Cambridge, MA).

Jongman, A. and Miller, J. D. (1991). "Method for the location of burst-onset spectra in the auditory-perceptual space: A study of place of articulation in voiceless stop consonants," J. Acoust. Soc. Am. **89**, 867–873.

Joos, M. (1948). "Acoustic phonetics," Language Suppl. **24**, 1–136. [Monogr. 23].

Just, M. A., Suslick, R. L., Michaels, S., and Shockey, L. (1978). "Acousic cues and psychological processes in the perception of natural stop consonants," Percept. Psychophys. **24**, 327–336.

Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (1999). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Commun. **27**, 187–207.

Kay, S. M. (1988). *Modern Spectral Estimation* (Prentice Hall, Englewood Cliffs, NJ).

Kewley-Port, D. (1982). "Measurement of formant transitions in naturally produced stop consonant-vowel syllables," J. Acoust. Soc. Am. **72**, 379–389.

Kewley-Port, D. (1983). "Time-varying features as correlates of place of articulation in stop consonants," J. Acoust. Soc. Am. **73**, 322–335.

Kewley-Port, D. and Luce, P. A. (1984). "Time-varying features of initial stop consonants in auditory running spectra: A first report," Percept. Psychophys. **35**, 353–360.

Kewley-Port, D., Pisoni, D. B., and Studdert-Kennedy, M. (1983). "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," J. Acoust. Soc. Am. **73**, 1779–1793.

Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.

Kopec, G. E. (1986). "Formant tracking using hidden Markov models and vector quantization," IEEE Trans. Acoust. Speech Signal Process. **ASSP-34**, 709–729.

Krull, D. (1990). "Relating acoustic properties to perceptual responses: A study of Swedish voiced stops," J. Acoust. Soc. Am. **88**, 2557–2570.

Lahiri, A., Gewirth, L., and Blumstein, S. E. (1984). "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study," J. Acoust. Soc. Am. **76**, 391–404.

LaRiviere, C., Winitz, H., and Herriman, E. (1975). "Vocalic transitions in the perception of voiceless initial stops," J. Acoust. Soc. Am. 57, 470–475.

Lehiste, I. and Peterson, G. E. (1961). "Transitions, glides and diphthongs," J. Acoust. Soc. Am. 33, 268–277.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). "Perception of the speech code," Psychol. Rev. 74, 431–461.

Liberman, A. M., Delattre, P., and Cooper, F. S. (1952). "The rôle of selected stimulus-variables in the perception of the unvoiced stop consonants," Am. J. Psychol. 65, 497–516.

Liberman, A. M., Delattre, P. C., and Cooper, F. S. (1958). "Some cues for the distinction between voiced and voiceless stops in initial position," Lang. Speech 1, 153–167.

Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. (1954). "The role of consonant-vowel transitions in the perception of the stop and nasal consonants," Psychol. Monogr. 68(8), 1–13.

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). "The discrimination of speech sounds within and across phoneme boundaries," J. Exp. Psychol. 54, 358–368.

Liberman, A. M. and Mattingly, I. G. (1985). "The motor thoery of speech perception revised," Cognition 21, 1–36.

Lindblom, B. (1963). "On vowel reduction," Tech. Rep. 29, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden.

Lindblom, B. (1990). "Explaining phonetic variation: A sketch of the H&H theory," in *Speech Production and Speech Modelling*, edited by W. J. Hardcastle and A. Marchal (Kluwer, Dordrecht, the Netherlands), pp. 403–439.

Lindholm, J. M., Dorman, M., Taylor, B. E., and Hannley, M. T. (1988). "Stimulus factors influencing the identification of voiced stop consonants by normal-hearing and hearing-impaired adults," J. Acoust. Soc. Am. 83, 1608–1614.

Lindsey, J. K. (1999). *Models for Repeated Measurements* (Oxford University Press), 2nd ed.

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

Lisker, L. and Abramson, A. S. (1967). "Some effects of context on voice onset time in English stops," Lang. Speech 10, 1–28.

Malécot, A. (1958). "The role of releases in the identification of released final stops," Language 34, 370–380.

Markel, J. D. and Gray, Jr., A. H. (1976). *Linear Prediction of Speech* (Springer-Verlag, Berlin).

Matthews, M. V., Miller, J. E., and David, Jr., E. E. (**1961**). "Pitch synchronous analysis of voiced sounds," J. Acoust. Soc. Am. **33**, 179–186.

McCandless, S. S. (**1974**). "An algorithm for automatic formant extraction using linear prediction spectra," IEEE Trans. Acoust. Speech Signal Process. **ASSP-22**, 135–141.

McCullagh, P. and Nelder, J. A. (**1989**). *Generalized Linear Models* (Chapman & Hall, London), 2nd ed.

Mullennix, J. W., Pisoni, D. B., and Martin, C. S. (**1989**). "Some effects of talker variability on spoken word recognition," J. Acoust. Soc. Am. **85**, 365–378.

Nearey, T. M. (**1992a**). "Applications of generalized linear modeling to vowel data," in *ICSLP '92 Procedings*, edited by J. J. Ohala *et al.*, International Conference on Spoken Language Processing (University of Alberta, Edmonton, Canada), pp. 583–586.

Nearey, T. M. (**1992b**). "Context effects in a double-weak theory of speech percpetion," Lang. Speech **35**, 153–172.

Nearey, T. M. (**1992c**). "Perceptually motivated models for speaker-independent recognition of stop+vowel syllables in English: Final report to AGT," Tech. rep., University of Alberta.

Nearey, T. M. (**1995**). "A double-weak view of trading relations: Comments on Kingston and Diehl," in *Phonology and Phonetic Evidence*, edited by B. Connell and A. Arvaniti, no. 4 in Papers in Laboratory Phonology (Cambridge University Press), pp. 28–40.

Nearey, T. M. (**1997**). "Speech perception as pattern recognition," J. Acoust. Soc. Am. **101**, 3241–3254.

Nearey, T. M. and Assmann, P. F. (**1986**). "Modeling the role of inherent spectral change in vowel identification," J. Acoust. Soc. Am. **80**, 1297–1308.

Nearey, T. M. and Shammass, S. E. (**1987**). "Formant transitions as partly distinctive invariant properties in the identification of voiced stops," Can. Acoust. **15**, 17–24.

Nossair, Z. B. and Zahorian, S. A. (**1991**). "Dynamic spectral shape features as acoustic correlates for initial stop consonants," J. Acoust. Soc. Am. **89**, 2978–2991.

Oden, G. C. and Massaro, D. W. (**1978**). "Integration of featural information in speech perception," Psychol. Rev. **85**, 172–191.

Ohde, R. N. and Sharf, D. J. (**1977**). "Order effect of acoustic segments of VC and CV syllables on stop and vowel identification," J. Speech Hear. Res. **20**, 543–554.

Ohde, R. N. and Sharf, D. J. (**1981**). "Stop identification from vocalic transition plus vowel segments of CV and VC syllables; A follow-up study," J. Acoust. Soc. Am. **69**, 297–300.

Ohde, R. N. and Stevens, K. N. (**1983**). "Effect of burst amplitude on the perception of stop consonant place of articulation," J. Acoust. Soc. Am. **74**, 706–714.

Öhman, S. E. G. (1966). "Coarticulation in VCV utterances: Spectrographic measurements," J. Acoust. Soc. Am. 39, 151–168.

Oppenheim, A. V. and Schafer, R. W. (1989). *Discrete-Time Signal Processing* (Prentice Hall, Englewood Cliffs, NJ).

Ostreicher, H. J. and Sharf, D. J. (1976). "Effects of coarticulation on the identification of deleted consonant and vowel sounds," J. Phonetics 4, 285–301.

Peterson, G. E. and Barney, H. L. (1952). "Cointrol methods used in a study of the vowels," J. Acoust. Soc. Am. 24, 175–184.

Pols, L. C. W. (1979). "Coarticulation and the identification of initial and final plosives," in *Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America*, edited by J. J. Wolf and D. H. Klatt (Acoustical Society of America, New York), pp. 459–462.

Pols, L. C. W. and Schouten, M. E. H. (1978). "Identification of deleted consonants," J. Acoust. Soc. Am. 64, 1333–1337.

Potamianos, A. and Maragos, P. (1996). "Speech formant frequency and bandwidth tracking using multiband energy demodulation," J. Acoust. Soc. Am. 99, 3795–3806.

Potter, R. K., Kopp, G. A., and Green, H. C. (1947). *Visible Speech* (D. Van Nostrand, New York).

Rabiner, L. R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE 77, 257–286.

Repp, B. H. (1984). "Closure duration and release burst amplitude cues to stop consonant manner and place of articulation," Lang. Speech 27, 245–254.

Repp, B. H. and Lin, H.-B. (1989). "Acoustic properties and perception of stop consonant release transients," J. Acoust. Soc. Am. 85, 379–396.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks* (Cambridge University Press, Cambridge, United Kingdom).

Schatz, C. D. (1954). "The role of context in the perception of stops," Language 30, 47–56.

Schouten, M. E. H. and Pols, L. C. W. (1983). "Perception of plosive consonants," in *Sound Structures: Studies for Antonie Cohen*, edited by M. van den Broecke, V. van Heuven, and W. Zonneveld (Foris, Dordrecht, the Netherlands), pp. 227–243.

Searle, C. L., Jacobson, J. Z., and Kimberley, B. P. (1980). "Speech as patterns in the 3-space of time and frequency," in *Perception and Production of Fluent Speech*, edited by R. A. Cole (Lawrence Erlbaum, Hillsdale, NJ), pp. 73–102.

Shammass, S. E. (1985). "Formant transitions, spectral shape, and vowel context in the perception of voiced stops," Doctoral thesis, University of Alberta.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," Science 270, 303–304.

Shao, J. (**1993**). "Linear model selection by cross-validation," J. Am. Statist. Assoc. **88**, 486–493.

Shao, J. (**1996**). "Bootstrap model selection," J. Am. Statist. Assoc. **91**, 655–665.

Sharf, D. J. and Hemeyer, T. (**1972**). "Identification of place of consoanant articulation from vowel formant transitions," J. Acoust. Soc. Am. **51**, 652–658.

Slaney, M. (**1994**). "Auditory toolbox," Tech. Rep. 45, Apple Computer, Inc.

Smits, R. (**1994**). "Accuracy of quasistationary analysis of highly dynamic speech signals," J. Acoust. Soc. Am. **96**, 3401–3415.

Smits, R., ten Bosch, L., and Collier, R. (**1996a**). "Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. I. Perception experiment," J. Acoust. Soc. Am. **100**, 3852–3864.

Smits, R., ten Bosch, L., and Collier, R. (**1996b**). "Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. II. Modeling and evaluation," J. Acoust. Soc. Am. **100**, 3865–3881.

Stevens, K. N. (**1971**). "Airflow and turbulence noise for fricative and stop consonants," J. Acoust. Soc. Am. **50**, 1180–1192.

Stevens, K. N. (**1975**). "The potential role of property detectors in the perception of consonants," in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. A. A. Tatham (Academic Press, London), pp. 303–330.

Stevens, K. N. (**1993**). "Models for the production and acoustics of stop consonants," Speech Commun. **13**, 367–375.

Stevens, K. N. and Blumstein, S. E. (**1978**). "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. **64**, 1358–1368.

Stevens, K. N. and House, A. S. (**1956**). "Studies of formant transitions using a vocal tract analog," J. Acoust. Soc. Am. **28**, 578–585.

Stevens, K. N., Manuel, S. Y., and Matthies, M. (**1999**). "Revisiting place of articulation measures for stop consonants: Implications for models of consonant production," in *Proceedings of the 14th International Congress of Phonetic Sciences*, vol. 2, pp. 1117–1120.

Summers, I. R. (**1991**). "Electronically simulated hearing loss and the perception of degraded speech," in *Bioinstrumentation and Biosensors*, edited by D. L. Wise (Marcel Dekker, New York), pp. 589–610.

Suomi, K. (**1985**). "The vowel-dependence of gross spectral cues to place of articulation of stop consonants in CV syllables," J. Phonetics **13**, 267–285.

Sussman, H. M. (**1991**). "The representation of stop consonants in three-dimensional acoustic space," Phonetica **48**, 18–31.

Sussman, H. M., Fruchter, D., and Cable, A. (**1995**). "Locus equations derived from compensatory articulation," J. Acoust. Soc. Am. **97**, 3112–3124.

Sussman, H. M., Fruchter, D., Hilbert, J., and Sirosh, J. (1998). "Linear correlates in the speech signal: The orderly output constraint," Behav. Brain Sci. **21**, 241–299.

Sussman, H. M., McCaffrey, H. A., and Matthews, S. A. (1991). "An investigation of locus equations as a source of relational invariance for stop place of categorization," J. Acoust. Soc. Am. **90**, 1309–1325.

Talkin, D. (1987). "Speech formant trajectory estimation using dynamic programming with modulated transition costs," J. Acoust. Soc. Am. Suppl. 1 **82**, S55.

Tekieli, M. E. and Cullinan, W. L. (1979). "The perception of temporally segmented vowels and consonant-vowel syllables," J. Speech Hear. Res. **22**, 103–121.

ter Keurs, M., Festen, J. M., and Plomp, R. (1992). "Effect of spectral envelope smearing on speech reception. I," J. Acoust. Soc. Am. **91**, 2872–2880.

Venables, W. N. and Ripley, B. D. (1998). *Modern Applied Statistics with S-Plus* (Springer, New York), 2nd ed.

Walley, A. C. and Carrell, T. D. (1983). "Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants," J. Acoust. Soc. Am. **73**, 1011–1022.

Warren, R. M. and Bashford, Jr., J. A. (1999). "Intelligibility of 1/3-octave speech: Greater contribution of frequencies outside than inside the nominal passband," J. Acoust. Soc. Am. **106**, L47–L52. [Published online 30 September 1999].

Winitz, H., Scheib, M. E., and Reeds, J. A. (1972). "Identification of stops and vowels for the burst portion of /p,t,k/ isolated from conversational speech," J. Acoust. Soc. Am. **51**, 1309–1317.

Zue, V. W. (1976). "Acoustic characteristics of stop consonants: A controlled study," Doctoral thesis, Massachusetts Institute of Technology.

Zwicker, E., Flottorp, G., and Stevens, S. S. (1957). "Critical bandwidth in loudness summation," J. Acoust. Soc. Am. **29**, 548–557.