Title
Acoustic Cues in the Directionality of Stop Consonant Confusions

Author
Plauché, Madelaine

Publication Date
2001

Acoustic cues in the directionality of stop consonant confusions

by

Madelaine Claire Plauché

B.S. (University of Texas, Austin) 1995
M.A. (University of California, Berkeley) 1997

A dissertation submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy
in
Linguistics

in the
GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor John Ohala, Chair
Professor Ian Maddieson
Doctor Elizabeth Shriberg
Professor Nelson Morgan

Fall 2001

Acoustic cues in the directionality of stop consonant confusions

by

Madelaine Claire Plauché

Abstract

Acoustic cues in the directionality of stop consonant confusions

by

Madelaine Claire Plauché

Doctor of Philosophy in Linguistics

University of California, Berkeley

Professor John J. Ohala, Chair

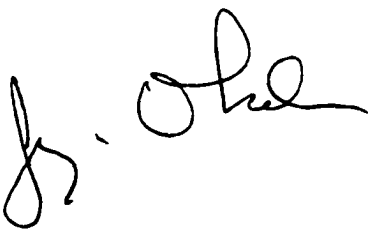
Stop place identification studies of consonant vowel sequences (CVs) in English (and many other languages) report similar patterns: Listener errors vary by (1) stop place, (2) the following vowel, and (3) confusion direction. *Asymmetries* in the direction of stop place confusions (e.g., [ki] is often confused for [ti], but [ti] is rarely confused for [ki]) have been noted by many researchers, but not yet adequately explained. This perceptual study evaluates previous accounts for asymmetry in stop place confusions and tests a new hypothesis: that listeners may disfavor stop places with greater token-to-token variation along a primary cue over those with less variation.

The current study differs in methodology from classical perceptual studies, in which generally one or two acoustic properties in synthetic or natural speech known to cue stop place are carefully manipulated. Instead, a combination of digital signal processing and machine learning techniques are used to examine a larger set of stop place cues to determine how they are ranked as well as how they vary by context and stop place. Additionally, the current study uses the natural token-to-token variation of speech

to avoid any artifacts that artificial or manipulated speech may cause in listener perceptions.

Unaltered CV tokens (C={p, t, k} and V = {i, a, u}) were collected from careful speech of native English speakers. Several potential cues to stop place were extracted for each token, including the rate and transition of the second formant during vowel onset, the gross spectral shape and relative amplitude of the stop burst, and voice onset time (VOT). CV tokens representing all nine contexts were chosen to serve as stimuli in a perception experiment. The unaltered stimuli were presented to experimental subjects for stop place identification in a cross-modal task designed to induce errors.

CVs that are more frequent in a listener's lexicon were more frequently given as incorrect responses in the stop place identification task. Additionally, perceptual asymmetry was caused by differences in token-to-token variation. In certain CV contexts, stimuli with ambiguous values were significantly more likely to be confused than those with canonical values. Phonotactic frequency in the lexicon, degradation of non-robust cues, and differences in production variation all affect perceptual error rates differentially by CV context. The results show that humans employ whatever cues (phonetic or non-phonetic) are available for the categorization of speech sounds, especially when primary acoustic cues are ambiguous.



Contents

Acknowledgments

I am greatly appreciative of my advisor, John J. Ohala, for showing me the importance of creativity in scientific pursuits, as well as to my committee members Ian Maddieson and Nelson Morgan, for their comments and support. Elizabeth Shriberg was a godsend early on in this project when she encouraged me to dabble in large amounts of data. It was a pleasure to conduct research in the stimulating (but quiet) atmosphere of the International Computer Science Institute; I will especially miss tea time. Thanks also to Lily Liaw, Julie Lewis, Alan Yu, and my other colleagues in the Phonology Lab, who contributed significantly to this work through in-depth discussions at various stages.

I was fortunate to have several talented technical people within earshot throughout this project. Thanks especially to Kemal Sönmez, Ron Sprouse, Andreas Stolcke, Dan Gildea, Scott Rodenheizer, Marina Ashiotou, the ICSI support staff and, of course, my live-in Windows and Linux networking specialist, Richard Carlson.

Thanks also to Nancy Chang for her companionship and all-night editing service. And Benjamin Bergen, thank you for kicking my derrière all the way through grad school. Okay, you can stop now. No really, that's enough. Sometimes it hurts to be your friend.

I am most grateful to Ashlee Bailey, Steve Chang, Dan Garcia, Mo Corston-Oliver and all of my Berkeley friends for our intense extracurricular schedule these last few years. You kept me fit, sane, and constantly amused.

Most importantly, I wish to thank my parents and my brothers for their love and encouragement. How quickly they learned not to ask: "Are you done yet?" I am especially fortunate for my partner, Richard Carlson. I owe the completion of this project to his dedication and support and did I mention that he fed me regularly? Thank you for each of our lovely mornings.

This research was funded by the National Science Foundation grant BCS #9817243 to the University of California, Berkeley.

To Wild Bill.

Chapter 1

Human perception and misperceptions

1 Introduction

The research presented in this thesis addresses the role of acoustic signal properties in the perception of place in the English voiceless, unaspirated, pre-vocalic stop consonants, [p, t, k]. Human perception of speech, and human perception of stop place in particular, has been the topic of phonetic research for many years. As a result, a vast literature is available on both the nature of cues to stop place and the nature of human errors in stop place categorization. Previous stop place identification studies of consonant vowel sequences (CVs) in English (and many other languages) report similar patterns: Listener errors vary by (1) stop place (alveolars are rarely confused), (2) the following vowel (high front vowels cause high confusion rates), and (3) confusion *direction*. *Asymmetries* in the direction of stop place confusions (e.g., [ki] is often confused for [ti], but [ti] is rarely confused for [ki]) have been noted by many researchers, but not yet adequately explained. The current study differs in methodology from classical perceptual studies, in which generally one or two acoustic properties in synthetic or natural speech known to cue stop place are carefully manipulated. Instead, a larger set of stop place cues are examined to determine how they are ranked as well as how they vary by context and stop place. The large set of cues permits a first attempt at explaining the directionality of stop place confusions. Additionally, the current study will use the natural token-to-token

variation of speech to avoid any artifacts that artificial or manipulated speech may cause in listener perceptions.

The first chapter summarizes previous findings in the domains of stop perception, specific cues to stop place, human errors in stop categorization, and parallels to attested historical sound changes in stop place. Section 2 presents the overall problem of stop consonant perception, beginning with a brief overview of the production of voiceless, unaspirated, stop consonants. Then a general model (synthesized from previous literature) of the stages involved in human perception of speech is provided in Section 3. Section 4 of this chapter presents the cues that humans are thought to use to detect place in voiceless, unaspirated stop consonants in CV position, as determined by the literature of perceptual studies. Section 5 reviews the findings of previous confusion studies. Cross-linguistic historical sound changes involving shifts in voiceless unaspirated stop place are summarized in Section 6, with a focus on stop place shifts that mirror perceptual confusions found in controlled laboratory studies. Two theories of sound change, both linking historical sound change to results of laboratory confusion studies, are also presented. Finally, Section 7 defines the specific objectives of the current research, including the investigation of the relative roles of acoustic cues in stop consonant confusions, specifically in their directionality.

2 The problem of stop consonant perception

The problem with stop consonant perception, and of speech perception in general, is the lack of a one-to-one correspondence between the acoustic signal and the linguistic categories that humans perceive. The production of a single linguistic unit (e.g.,

phoneme, distinctive feature, syllable, morpheme) is subject to considerable acoustic variation due to phonetic context, stress, physiology of the speaker (vocal tract length, vocal cord length, etc.), sociological characteristics of the speaker (dialect, gender, rate of speech, relationship with interlocutor, etc.), token-to-token variation, as well as the environment of the utterance (room acoustics, reverberation, background noise, etc.).

One of the primary contributions to variability is the overlapping nature of the speech signal (Liberman et al. 1967; Fant 1973; Lindblom 1986). The speech signal cannot be separated into discrete phonetic units, but instead is the product of *coarticulation*, the continuous motion of articulators from one phonetic target to the next, resulting in phonemes that are highly variable across contexts (Öhman 1966). There is no direct correspondence between the acoustics of the speech signal and the phonemes intended by the speaker. The fact that humans can identify phonemes from continuous speech with the accuracy they do given the large degree of variation and the continuous, overlapping property of the speech signal has puzzled researchers for over 50 years.

The stages of production of stop consonants and the acoustic consequences of each stage are well known. Stop consonants are a class of phonemes characterized by their production, which involves a full constriction in the vocal tract by an articulator (usually the lips or the tongue). The current research focuses on the voiceless, unaspirated stop consonants of English [p, t, k] that occur primarily in [s]-initial clusters. Part of the reason that unaspirated, voiceless stops were chosen in this study is that although they are somewhat phonologically restricted in English, they are included in the inventories of 80% or more of the world's languages sampled by UPSID (Maddieson 1984), an archive

that contains data on 317 languages chosen to sample the different language families of the world.

During the production of a stop consonant in #[s]CV position, the articulator moves from a position of less constriction to create a full closure at some site in the oral cavity. Generally the full constriction lasts no longer than 100 msec, during which time the air pressure upstream from the constriction builds, creating a burst when the closure is released. The stages in the production of all voiceless unaspirated stop consonants in #[s]CV position and the acoustic consequences of each stage are summarized in Table 1.

Articulation of a Stop Consonant	Corresponding Acoustic Events
The articulator moves toward a full constriction.	The amplitude of the signal decreases to its minimum.
*A **full constriction*** is caused by the articulator. Pressure behind the closure builds rapidly, possibly causing a passive expansion of the oral cavity by compliance of the walls.	The ***closure*** is a period of zero amplitude.
During the ***release***, the articulator moves away from the constriction and toward the following vowel configuration, causing an abrupt increase in airflow at the partially constricted area.	The ***burst***, ***release***, or ***transient*** is the initial burst of noise and abrupt increase of amplitude with particular spectral characteristics.
As airflow continues across the constrictions of the consonant and the following vowel, and through the glottis, there is a rapid decrease in intraoral pressure, causing the walls of the vocal tract to return to their rest position (i.e., *recoil*).	***Frication*** is turbulence in the airflow that can occur at the partially constricted area subsequent to or overlapped with the release.
In the case of phonetically aspirated stops, the vocal cords remain open before they are adducted for the following vowel.	In the case of phonetically aspirated stops, turbulence or ***aspiration*** noise is generated in the airflow across the glottis. Formant structure in low frequencies may emerge.
Vocal fold vibration for the following vowel begins when the supraglottal pressure decreases and the vocal folds are adducted.	At the ***voicing onset*** of the following vowel, a periodic signal with vowel formant structure emerges.
The articulator continues movement toward configurations appropriate for the following vowel.	***Formant transitions*** are the result of changes in resonant frequencies created by the articulator as it moves toward the vowel configuration.

Table 1. Stages of #[s]CV consonant production and the corresponding acoustic consequences (Fant 1973; Stevens 1999).

Cues for the *place of articulation* of stop consonants lie in each of the acoustic events mentioned in Table 1. Stop consonants in English (and many other languages) are produced at three main places of articulation (Figure 1). A *bilabial* stop (e.g., English /p, b/) is produced by a complete closure at the lips. An *alveolar* stop (e.g., English /t, d/) is formed when the apex of the tongue is pressed against the alveolar ridge. When the dorsum of the tongue is pressed against the palate, a *velar* or *dorsal* stop (e.g., English /k, g/) is produced.

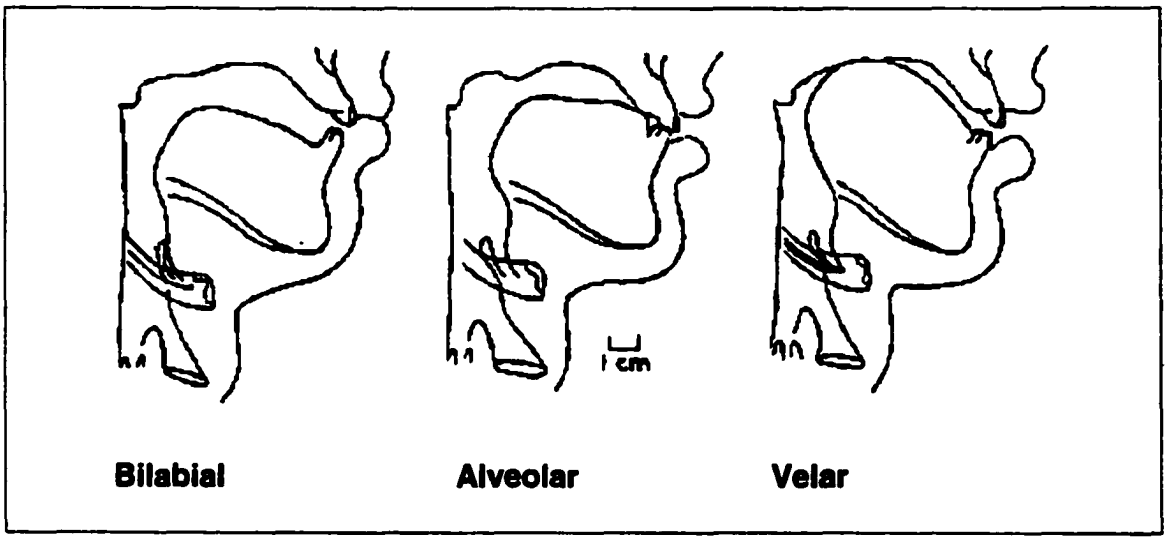


Figure 1. Mid-sagittal diagrams of the vocal tract during the closure of bilabial, alveolar, and velar stops (from Perkell 1969).

The production of stops at all three places of articulation share the general sequence of events and corresponding acoustic consequences listed in Table 1. Each of these acoustic events, however, assumes different spectral and temporal characteristics depending on the place of the articulation of the closure. In Section 4, the particular articulatory and acoustic characteristics of stop production are examined for each of the three places of articulation: bilabial, alveolar, and velar.

As with phonemes in general, all of the acoustic events associated with stop consonant production are highly variable depending on phonetic context, especially the following vocalic context. Also, the abrupt shifts in the vocal tract configuration during stop consonant production result in an acoustic signal that changes rapidly over time, especially in the spectral domain. Somehow, humans are able to extract from this signal the necessary information for identifying the phoneme produced, presumably by integrating several perceptually relevant acoustic properties at a time.

3 Human stop perception

This section synthesizes assumptions and findings in the literature on the perception of phonemes and stop consonants (cf. Smits 1995 for a summary). During perception, the speech signal generated by the speaker must first pass through the surrounding environment and through the auditory system of the listener. Both the environment and the human auditory system are known to cause various transformations to the acoustic signal (Gold & Morgan 2000). Speech sounds are transmitted via the outer and middle ear to the basilar membrane, which acts as a filter bank, exciting auditory nerves that are tuned to specific frequency bands via their position along the basilar membrane. When a stimulus is applied with an abrupt onset of amplitude, certain auditory nerve fibers fire at a higher rate (more spikes per second). If the stimulus has a more gradual onset of amplitude, the rate of nerve firing decreases but remains steady. This property of auditory nerves suggests that the auditory system is highly sensitive to dynamic cues (Delgutte 1980). The ear is also known to be most sensitive to frequencies between 100 Hz and 10,000 KHz, effectively band-pass filtering the speech signal into bands whose width

increases proportionally with the frequency. The transformation of the speech signal by the human auditory system may be simulated by presenting speech in perception studies up to 8 KHz only, with a *pre-emphasis* to the signal, or a boost to the higher frequencies of the signal. Though there has been much research on the characteristics of the human auditory system, the exact effect of this system on an acoustic speech signal is unknown. For this reason, it is difficult to estimate the perceptual weight that the human auditory system will attribute to an acoustic property of a signal. In the current research, very little effort is made to artificially simulate the human auditory system. Instead, with the exception of pre-emphasis to the higher frequencies of the signal, we rely entirely on the acoustic properties of the signal in estimating cues. The acoustic characteristics of the signal examined in the current thesis could potentially be adjusted to simulate the effect of the auditory system in further studies.

The perceptual system is also thought to be equipped with an *event-detector*, a mechanism that detects periods of abrupt spectral change, such as stop closures and bursts, and voicing onset (Blumstein & Stevens 1981; Stevens 1999). Once an acoustic event such as a stop closure or burst is detected, the perceptual system can pinpoint the portions of the signal that provide the appropriate information about the nature of the phoneme. In the current study, the burst release and voicing onset were labeled by hand by the author, and from these landmarks in the signal, acoustic properties thought to be cues to stop place were automatically extracted. Not only does this method partially imitate the event detection described, but it is also particularly applicable to machine recognition. There is a growing effort to develop data-driven approaches to front-end adaptation in Automatic Speech Recognition (ASR) (Weinstein et al. 1975; Cole et al.

1986; Glass et al. 1996; Bacchiani & Ostendorf 1998; Sönmez et al. 2000). These approaches often rely on automatic event detection in combination with the extraction of acoustic properties known to cue phoneme identity.

During stop consonant perception, listeners are also known to extract and adapt to background information, such as speaker characteristics (Mullennix & Pisoni 1990), noise (Miller & Nicely 1955), filtering (Shriberg 1992), and speaking rate (Miller & Sachs 1981). In the current study, all of these conditions, with the exception of speaker characteristics, are held as constant as possible.

In order to categorize speech, the human perception system is thought to combine information from the event detector, background information, and *top-down* information that includes both linguistic context of all levels (phonological, morphological, etc.) and non-linguistic context (social, knowledge of the speaker, knowledge of the world, etc.). The combined information at this level determines what relevant acoustic cues are to be extracted. The values of these cues are then integrated into a linguistic classifier, whose proposed linguistic units include: distinctive features, allophones, segments, and syllables (Wickelgren 1976). The precise mechanism for the classification of linguistic categories from speech is unknown. Some have proposed that classification is based on prototypes (Kuhl 1992), while others believe the human classification is more sensitive to categorical boundaries (cf. Lahiri et al. 1984). The current study allows for either a prototype or category boundary type of stop place classification.

4 Review of cues to place in stop consonants

We have discussed the general mechanisms thought to be involved in the human perception of linguistic units. In this section, we delve more specifically into cue extraction by reviewing the literature of cues to stop place. Although the majority of the reviewed studies were conducted on voiced stops or initial voiceless stops, many of the cues to place are shared by voiceless, unaspirated, pre-vocalic stops, the focus of the current research. Cues to stop place for stop consonants in VC or VCV contexts, such as closure duration (Repp 1984a, 1984b), were not included in this study.

A property of the acoustic signal is usually considered a cue if excising a portion of the signal or manipulating a property of the signal affects listeners' responses. In many studies investigating the cues to stop place, the background information and speaker characteristics as well as token-to-token variation were controlled by the use of synthetic stimuli. In the current study, background information is controlled as much as possible, but the natural speech of multiple speakers is used for analysis and as stimuli in the perception study, to support claims about naturally occurring ambiguity in the signal. Of interest to this study are variations in specific cues to place of articulation in voiceless stop consonants in #[s]CV context and the relative informativeness of these cues depending on the following vowel context.

4.1 The burst release

The release of the burst in CV contexts carries a great deal of perceptual information. When the burst is excised from speech and the remaining transitions and following vowel are played to listeners, errors in place of articulation identification rise significantly

(Dorman et al. 1977; Ohde & Sharf 1977; Repp & Lin 1989). On the other hand, when only the initial 25 msec (including the burst and initial transitions) of synthesized CV stops are presented to listeners, they are still able to identify stop place consistently (Stevens & Blumstein 1978). Place of articulation information of the stop consonant is thought to lie in the duration, amplitude, and gross spectral shape of the burst and following frication.

Duration of the burst release

Researchers agree that the duration of the burst release correlates with the place of articulation in the following order: [p] < [t] < [k] (Dorman et al. 1977; Zue 1976). Due to difficulties in automatically determining the end of a burst release from a spectrogram, the duration of burst release was not directly extracted in this study, though aspects of this property are captured by VOT (Section 4.3) and the occasional presence of multiple bursts (Section 4.4).

Amplitude of the burst release

The amplitude of the burst release also cues listeners to place of articulation for voiceless unaspirated pre-vocalic stops. Ohde & Stevens (1983) found that lowering the relative amplitude of the high frequency component of a C[a] burst resulted in more [p] ratings than [t] ratings. Repp (1984b) found that burst amplitude was a cue for both stop manner and place of articulation. In a perception study in which he low-pass filtered stop consonants [p] and [t] at 4.8 KHz and manipulated the amplitudes of their bursts by 10

dB, the higher burst amplitudes caused 't' responses and lower amplitudes favored 'p' responses. Hedrick & Jesteadt (1996) found similar results with some variation depending on the following vowel of CV segments. When the overall relative amplitude of the burst was increased with respect to the vowel, [p]'s preceding the vowel [a] or [i] caused fewer 'p' responses. The rates of 'p' responses for [p]'s preceding [u], however, were less affected by the relative amplitude of the burst. The effect of increased misidentification of [p]'s with increased burst amplitude was confirmed for Spanish stop consonants preceding the high front vowel [i] by Plauché et al. (1997), who found that confusions between [pi] and [ti] increased when the burst amplitude for [pi] was doubled with respect to the vowel.

The amplitude of the burst may also be a cue for distinguishing between bilabials and velars. Fischer-Jørgensen (1967) found that for Danish voiceless unaspirated stops, the strength of the burst was greater for [k] than for [p], except in front of rounded vowels where the amplitude of [k]'s was greater than [p]'s only 50% of the time.

Ohde & Stevens (1983), however, expect that the amplitude of the higher-frequency components is more important for stop manner perception than for place of articulation perception. In the current study, amplitude of the burst is measured across all frequencies, though the set of spectral properties at the burst is sensitive to differences in distribution of energy over the burst spectrum.

Spectral characteristics of the burst release

The center frequency of the burst release is known to cue stop place for pre-vocalic voiceless stops. Pattern playback studies (Cooper et al. 1952) that varied the center

frequency of synthesized bursts demonstrated that synthesized CV tokens with high-frequency bursts were heard as [t] across all vowel contexts, whereas both [k] and [p] judgments varied depending on the following vowel. Bursts at lower frequencies caused 'k' responses if they were slightly above or level with the formant of the following vowel; otherwise, listeners heard 'p'. Repp & Lin (1989) noted that in the burst onset spectra of bilabials in all the vowel contexts examined, there was a prominent dip around 1000 Hz that was absent for both /d/ and /g/ in the same contexts.

Additional evidence for the importance of the high-frequency spectral information for perception in the auditory domain was found by Delgutte (1980), who showed that certain auditory-nerve fibers fired rapidly in response to speech-like noise bursts with abrupt onsets (as in the case of high-frequency alveolar bursts), but less rapidly in response to more gradual onsets (as for bilabial stops). This suggests that listeners are more likely to respond to high-frequency energy than overall energy.

Bilabial, coronal, and dorsal stops vary systematically in the spectral characteristics of the acoustic signal in the 10 to 20 msec following the stop release. At the release, Blumstein & Stevens (1979) found that for a spectrum taken with an 8 ms window from 0 to 5 KHz, bilabials had a diffuse-falling spectrum (high energy in the low frequencies), alveolars had a diffuse-rising spectrum (high energy in the high frequencies), and that velars had a compact spectrum (a prominent spectral peak in the mid-frequency range from 1 to 3 KHz, depending on the following vowel).

Several investigators have noted that the spectral characteristics of stop consonant bursts serve as a cue to place of articulation (Cooper et al. 1952; Winitz et al. 1972; Blumstein & Stevens 1978), especially when more reliable cues, such as formant

transitions, are obscured. However, not all researchers have found this result (cf. Repp 1984b). In particular, the importance of the mid-frequency peak as a cue for velarity was confirmed when Plauché et al. (1997) band-pass filtered the mid-frequencies in Spanish velar stop bursts occurring in front of [i] and played the segments to listeners. The subjects gave 't' responses at significantly higher rates.

It has more recently been suggested that dynamic cues are the primary cues in stop place identification. In particular, the distinction between bilabial and alveolar stop consonants is proposed to rely on the manner in which the spectrum shape at the burst *changes* in the initial few tens of milliseconds following the consonantal release (Lahiri et al. 1984; Kewley-Port 1982, 1983; Stevens 1999). Indeed, much of the recent research has examined the role of dynamic cues, changes in formants or spectra over time that are expected to play an important role in stop place perception (cf. Smits 2000). Due to restrictions in the method of semi-automatic extraction in the current study, the majority of the cues explored here are static. Adapting the current methodology to an expanded set of cues including dynamic or more finely grained acoustic measures, however, would be a natural direction for further research.

4.2 Formant Transitions

Formant transitions, in particular second formant (F2) transitions, are considered primary cues for stop place (Cooper et al. 1952; Delattre et al. 1955; Liberman et al. 1967; Kewley-Port 1982, 1983). Formant transitions are also known to carry information about the vowel (Liberman et al. 1967), which explains their large amount of variation by vocalic context. In conflicting cue tasks, listeners presented with an onset spectrum

specifying one place of articulation spliced to formant transitions specifying another rely primarily on the information provided by the formant transitions to determine stop place (Dorman et al. 1977; Ohde & Sharf 1977; Dorman & Loizou 1996).

During the release of a stop consonant in CV position, the lips and tongue body move toward the target configuration for the production of the following vowel. The articulator movement causes changes in vocal tract resonances, resulting in formant transitions in the acoustic signal that vary by the place of articulation of the stop and by the following vocalic context. The opening of the lips during the release of a bilabial stop, for example, causes the formants to rise, with the first formant (F1) generally beginning to rise within the first 10 msec. Alveolars in CV contexts, however, may show a small fall or rise of F2 transitions, depending on the following vowel, as well as a slower rise of F1 as compared to bilabial stops. During a velar release into a following vowel, especially a back vowel, the second formant and third formant are often close together, forming the characteristic *velar pinch*. This characteristic varies greatly depending on the following vowel as well as the language studied (cf. M. Ohala 1995 on Hindi).

Results from pattern playback studies (Cooper et al. 1952; Delattre et al. 1955) in which the F2 and F3 formants of synthesized /b/, /d/, and /g/ were systematically varied suggested that place of articulation cues resided in extrapolated values of F2 and F3 at the burst release derived from the direction and rate of change of F2 and F3 in the following vowel. These extrapolated values, or *loci,* mapped to stop place categories. /b/ was characterized by a locus at 720 Hz and /d/ by a locus at 1800 Hz. The locus for /g/ at 3000 Hz could be found only for those cases in which the adjoining vowel had a second

formant above 1200 Hz. Formant transition variability by vocalic context was further studied in natural speech stimuli (Kewley-Port 1982). Locus equation studies built on the notion of *loci*, showing that F2 and F3 information, especially the regression from F2 values at vowel onset to values at the vowel midpoint, map to the three stop places fairly consistently across vocalic contexts (Sussman et al. 1991).

Additionally, the *rate* of the formant transitions (especially F1) varies according to the place of the stop articulation. Velar formant transitions are slower than those of corresponding bilabial and alveolar stops, due to the mass of the articulator involved (i.e., the body of the tongue) (Kewley-Port 1982; Stevens 1999).

Many phonetic studies investigated the place information that formant transitions convey in a single vocalic context, usually [a]. An exception to this is Ohde & Sharf (1977), who presented listeners with the burst and vocalic transition only and examined their responses to consonant place, [p,t, k], in CV position across the vowels [i, er, u]. Alveolars were identified equally well across all vocalic contexts. Rates of identification for [p] increased significantly from [i] to [er] to [u]. Velars were highly confused when they preceded [i] and [u], but fared well before [er].

4.3 VOT (Voice Onset Time)

Voice onset time (VOT) is the duration from the release of the burst to the onset of periodic vocal fold vibration. VOT is a primary cue cross-linguistically to the perception of stop manner, i.e., distinctions between voiced, voiceless, and aspirated stops (Lisker & Abramson 1967; Fant 1973). VOT is also known to increase progressively in duration as

the place of articulation moves back from bilabial to alveolar and then to velar within a given manner class (Lisker & Abramson 1967; Klatt 1975).

In a cross-contextual study on stops in [s]C clusters, Klatt (1975) found that the average VOT for voiceless, unaspirated stop consonants was 12 msec for bilabial, 23 msec for alveolar, and 30 msec for velar stops. VOT was also found to vary systematically by vowel height within voiced and voiceless stops (Lehiste 1970; Klatt 1975). Chang (1999) showed that the close oral constriction of the high vowels [i, u] creates greater impedance to the air escaping from the mouth, thus slowing the drop in oral pressure necessary for the initiation of voicing.

Variations in VOT by stop place (bilabial < alveolar < velar), present even in aspirated stops, have been proposed to be due to phase-locking (Maddieson 1997), as the stop *closure* durations are generally ranked in the opposite order (velar < alveolar < bilabial). The sum of the durations of the closure period and the duration of the transient plus aspiration (VOT) are approximately the same across all stop places, regardless of aspiration (Weismer 1980), suggesting that variations in VOT are a side effect of the fixed length for the abduction-adduction cycle of the vocal cords for voiceless stops.

Other researchers have proposed that variations in VOT by stop place and by vowel are due to the physical constraints necessary for voicing (Ohala 1981; Stevens 1999; Chang 1999). VOT is thought to depend on the rate of increase of the cross-sectional area of the constriction, since this dictates the rate of pressure shift from high oral pressure to a transglottal differential necessary for voicing. When the place of articulation is at the lips, the rate of opening is high, yielding a short VOT for this place of articulation. A velar release, on the other hand, involves a more massive articulator

(tongue body) and a release that is non-perpendicular to the constriction (Houde 1968; Keating & Lahiri 1993; Hoole et al. 1998), resulting in a slower rate of constriction opening. The slow opening demands a longer period of time for the appropriate transglottal pressure differential for voicing to be established. The alveolar stop is produced with the tongue tip nearly perpendicular to the palate. The rate of constriction opening and the correlated duration of VOT is somewhere between that of a labial stop and that of a velar stop (Fujimura 1961).

Regardless of the precise cause of variation in VOT by stop place, speakers appear to be aware of the association. Studies on the perception of VOT have shown that the thresholds of voiced-voiceless for English speakers is subject to highly significant main effects and interactions for both the effects of place of articulation and of vocalic context (Summerfield & Haggard 1977; Nearey & Rochet 1994).

4.4 Multiple bursts

Another characteristic of stop consonants, which may or may not be used by speakers to identify stop place, is the occasional presence of multiple burst releases (usually two or three distinct burst releases, and as many as five) in voiceless velar stops (Repp & Lin 1989), thought to result from the large area of the velar constriction and its relatively slow release (Stevens 1999). Multiple bursts can occur with coronals and even bilabials before closely rounded vowels (Repp & Lin 1989), but Stevens (1999) estimates that the phenomenon is more likely when the release is slower and the constriction is longer, as it is in the release of a velar or uvular constriction.

The production of a multiple burst is thought to be caused by the interactions of the intraoral pressure, which builds rapidly during the closure and then drops suddenly after the release (Stevens 1999). As the tongue body lowers slowly to initiate the release, the natural compliance of the articulator surface involved in the production of the full constriction is subject to the Bernouilli force caused by airflow through the still quite narrow constriction, which in effect sucks the tongue surface back toward the palate at quick regular intervals (Figure 2).

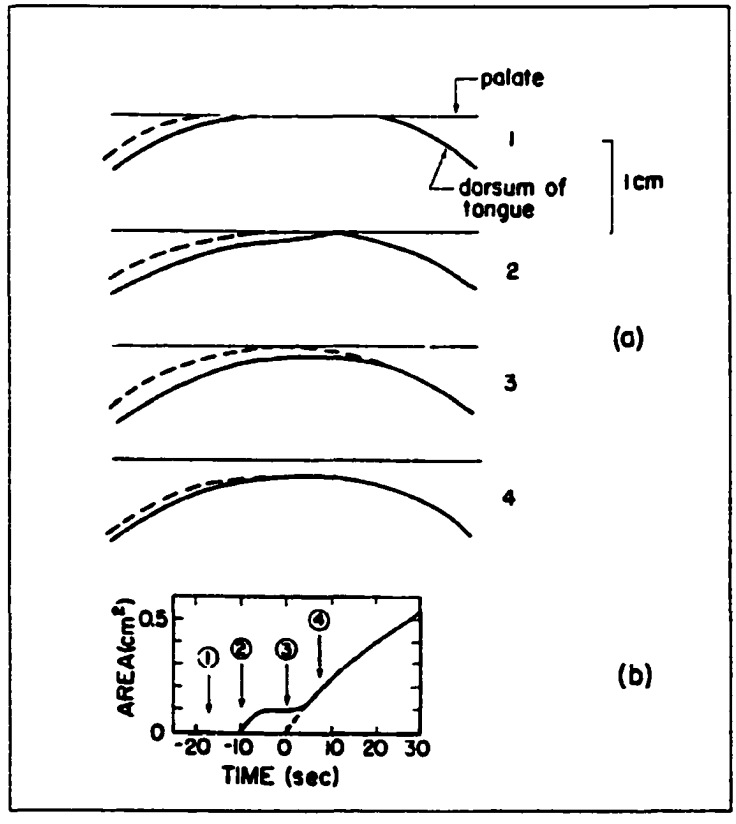


Figure 2. Schema of the area of constriction during a stop release (from Stevens 1999).

The role of multiple bursts in stop place perception is unknown at this time. Although critical band studies (Gold & Morgan 2000) and successive *click* studies, where listeners are presented with a series of clicks separated by varying intervals (Lehiste 1970), predict that the auditory system is unlikely to be able to separately resolve the successive rapid transients, multiple burst releases may indirectly reinforce the spectral information of the burst, which is itself rich with place of articulation information. The current study will investigate the role of multiple bursts in listeners' identification of stop place.

4.5 Summary of place of articulation cues

Table 2 is a summary of the expected values along the relevant cues for the three places of articulation in [s]CV context, as described by sections 4.1- 4.4. Consistent variations by the following vowel ([i, a, u]) are included, wherever possible.

Cues to stop place	Bilabial	Alveolar	Velar
Burst amplitude	Low	High	[u] - Low [i,a] - High
Gross spectral shape at burst	Falling	Rising	[a,u] - Compact [i] - Rising
Spectral center at burst	Low	High	[a,u] - Low [i] – Mid
F2 onset	[i] - High [a] - Mid [u] - Low	High	[i] - High [a] - Mid [u] - Low
F2 transition	Rising	[u,a] - Falling [i] - Rising	Falling
F3 onset	Low	High	Low
VOT	[a] - Short [i, u] - Short/Med	[a] - Med [i,u] - Med/Long	Long
Multiple bursts	[a] - Rare [i, u] - Rare/Occurs	[a,u] - Occurs [i] - Occurs/Common	Common

Table 2. Summary of place of articulation cues. Spectral shape values correspond to the shape of the overall spectrum at the burst: 'Falling' refers to a spectrum with the majority of the energy in the low frequencies, a 'Rising' value is given to spectra with a majority of energy in the high frequencies (up to 8 KHz), and a 'Compact' value is the label for bursts with the majority of energy in the mid-frequencies. For burst amplitude, the labels 'Low', 'Med', 'High' refer to the amount of energy. In the spectral peak of the burst and the onset of F2 and F3, however, the same labels refer to the frequency.

The cues presented in Table 2 are all predicted to be useful in the place identification of voiceless unaspirated stop consonants. One of the tasks of this study is to

examine the relative importance of these cues for determining stop place and how this ranking differs by context and by stop.

Previous conflicting cue studies that present listeners with manipulated stimuli in which one cue signals one place of articulation and a second cue signals a different place of articulation have made claims about the interacting effects between the two (or sometimes three) cues examined. For example, Dorman & Loizou (1996) showed that when formant information is relatively non-distinct between bilabials and alveolars, other information has the opportunity to influence identification. This suggests that formant transitions are primary for stop place information, while other cues play a role only when formant information is ambiguous. In a similar study, Fant (1973) found that for Swedish stops in front of the vowel [a], the F2 transitional cues are almost identical for [ta] and [ka] and that normally in this case, the burst provides a distinguishing cue. Other perceptual studies, however, showed that the release burst of velar stops, in particular, is more heavily weighted as a perceptual cue to place of articulation than are the following formant transitions (Dorman et al. 1977; Smits 1995).

The current study will continue the effort to rank cues for stop place by examining a larger set of stop place cues than traditional conflicting cue studies, which investigate at most two cues at a time. This study will also use the token-to-token variation of natural speech to avoid any artifacts that artificial or manipulated speech may cause in listener perceptions. Though the cues listed in Table 2 represent a subset of stop place cues, if we make the reasonable assumption that listeners have access to at least this set of cues in CV utterances, then the current study is a good first approach to solving the problem of context-dependent cue discriminibility. In the following section, we turn to a summary of

confusion studies, with a focus on recurring patterns of confusions in the identification of place of articulation of stop consonants. Recurring confusions across languages and experiments may indicate contexts in which important cues to stop place, such as formant transitions, are ambiguous, forcing listeners to rely on secondary cues to determine stop place.

5 Stop consonant confusions

So far, we have discussed the general mechanisms involved in human stop perception and the particular acoustic properties known to contain information about stop place for voiceless, unaspirated, pre-vocalic stop consonants in English. In this section, we present a number of stop confusion studies, with a particular focus on three important trends that these studies report: (1) Error rates vary by stop place (e.g., alveolar stops are less likely to be confused than velar or bilabial stops), (2) error rates vary by vocalic context (e.g., [i] causes the highest rate of preceding stop confusions), and (3) confusions may be asymmetric (where stop A is often confused for stop B but stop B is rarely confused for stop A, noted A→B). We also review some of the explanations for asymmetries in phoneme identification that have been suggested so far, including accounts based on knowledge of the following vowel, markedness, and frequency of the phoneme in the lexicon of the speaker/listener.

5.1 Review of stop confusion studies

Many previous confusion studies focused on the effect of noise (Miller & Nicely 1955; Wang & Bilger 1973), filtering (Miller & Nicely 1955), presentation level (Bell et al.

1989), language (Ahmed et al. 1969; Singh & Black 1966), normal versus hearing (Hedrick & Ohde 1996), and syllable position (Ahmed & Agrawal 1969; Bell et al. 1989) on listener error rates of near-natural stop consonants. The majority of these studies, however, examined stop consonant confusions in a single vowel context, usually [a], or reported experimental results summed across the vocalic contexts. In contrast, the current study examines how the following vowel context ([i, a, u]) affects error rates in the place identification of the previous stop.

Previous confusion studies unanimously report that English listener confusion rates are significantly lower for pre-vocalic alveolar stop consonants than for bilabial and velar stop consonants across all vowel contexts examined (Miller & Nicely 1955; Tekieli & Cullinan 1979). Additionally, alveolars in C[a] position were slightly better recalled in a short-term recall task than other stop places (Wickelgren 1966).

Velars stops in pre-vocalic position across all vowel contexts generally cause the highest confusion rates when presented to English listeners (Wang & Bilger 1973; Stevens & Blumstein 1979; Winitz et al. 1972; Repp & Lin 1989), though velars were more easily identified in VC contexts (Wang & Bilger 1973).

In Hindi, however, confusion rates summed across all 10 Hindi vowel contexts revealed that [t] was more often confused than [p] or [k] when presented as the first consonant in CVC syllables (Ahmed & Agrawal 1969). Instead, an initial retroflex [ʈ] caused the highest rates of confusion in Hindi CVC syllables.

In confusion studies in which portions of the signal are excised or altered, listeners' confusion rates in stop place identification are higher for pre-vocalic stop consonants, especially [p] and [k], when the vowel is a high front vowel (e.g., [i]) than

any other vocalic context (Singh & Black 1966; Wang & Bilger 1973; Stevens & Blumstein 1979; Winitz et al. 1972; Repp & Lin 1989; Neagu 1998). Confusion rates for alveolar stops in pre-vocalic position, however, show less variation across examined vocalic contexts (Winitz et al. 1972).

In one of the few confusion studies that compared rates across three vowel contexts, Wang & Bilger (1973) found that identification results for all consonants (including voiceless, voiced, fricatives, and affricates) in CV position increased from [i] to [a] to [u] when listeners were presented with the burst and 100 msec of the following vowel. In a similar experiment, [k]'s were more often misclassified as [p]'s when they preceded the rounded vowel [u] (Repp 1989).

Winitz et al. (1972) reports consonant confusions across the vowels [i, a, u] (Table 3). Stimuli of the stops [p, t, k] that included the stop burst and the following 100 msec were presented to listeners for identification in CV and VC position (thus the stops were presumably voiceless *aspirated*). The results from the initial consonant identification are shown in Table 3.

Spoken	[p]			[t]			[k]		
Heard	**[i]**	**[a]**	**[u]**	**[i]**	**[a]**	**[u]**	**[i]**	**[a]**	**[u]**
P	**.46**	**.83**	**.68**	.03	.15	.10	.15	.11	.24
T	.38	.07	.10	**.88**	**.63**	**.80**	.47	.20	.18
K	.17	.11	.23	.09	.22	.11	**.38***	**.70**	**.58**

Table 3. Confusion matrix from Winitz et al. (1972). Correct responses are shown in bold. Correct responses at less than chance are starred. Values do not sum to 1 due to rounding errors.

The alveolar [t] has the highest identification rates, though significant confusions occur when [t] is followed by [a]. The segment [ki] is identified at less than chance, where it was most often heard as a [ti]. Velars in other contexts are identified at rates comparable to bilabials. [pi] was often confused for [ti], though the reverse is not true. Bilabials and velars were often confused for each another in the context of [u].

In a comparable study in Italian, Delogu et al. (1995) examined the segmental intelligibility of Italian text-to-speech systems, using an open-response test of all phonemes known to the listeners as possible responses. VCV and CV sequences of meaningful and meaningless Italian words were used as stimuli, where the vowel portion consisted of [i], [a], or [u]. The results from the identification of the voiceless unaspirated stops [p, t, k] in CV position reveal results similar to English (Table 4).

Spoken	[p]			[t]			[k]		
Heard	**[i]**	**[a]**	**[u]**	**[i]**	**[a]**	**[u]**	**[i]**	**[a]**	**[u]**
P	**.19**	**.88**	**.31**	.13	0	0	.19	0	.06
T	.81	.06	.13	**.77**	**.94**	**.87**	.75	.13	.19
K	0	.06	.56	0	.06	.13	**.06**	**.77**	**.75**

Table 4. Italian stop confusion matrix (from Delogu, p.c.). Correct responses are shown in bold.

Though the rates differ from the Winitz et al. (1972) study, the robustness of alveolars and the vowel-dependent variations for [p] and [k] are similar. In this study, however, [pu] was often heard as *k*, but the reverse was not true. [pi] and [ki] were both confused for [ti] a majority of the time. [ti], however, was correctly identified 77% of the time.

Asymmetric confusions such as [ki]→[ti] and [pi]→[ti] have been noted by many researchers, but an adequate explanation for this phenomenon has yet to be offered. Here, we review some of the most common asymmetric confusions as well as the current explanations for this phenomenon.

5.2 Previous explanations for asymmetries in confusions

As we have seen in this chapter so far, velars that precede high front vowels are commonly misidentified as alveolars, but alveolars in the same context are not subject to the same confusion rates by listeners (e.g., [ki] → [ti]) (Winitz et al. 1972; Stevens &

Blumstein 1979). Bilabials that precede high front vowels are often mistaken for alveolars by listeners, but the reverse is not true (i.e., [pi]→[ti]). On the other hand, English velars and bilabials that precede [u] are confused for one another at comparable rates (Winitz et al. 1972; cf. Delogu (1995) for Italian stops).

Previous explanations for unidirectional confusions often rely on notions of *similarity* of the acoustic signal. Stevens & Blumstein (1979) explain the misclassification of labial and velar consonants as alveolars in the context of high front vowels [i] and [e] by the relatively high-frequency onset of F2 and F3 for both places of articulation in the environment of these vowels. In a study of Hungarian CVs, Blumstein (1986) examines the acoustic structure of voiceless velar and palatal stops in the environment of [i] and [u]. She analyzes the shift from [ki] to [ci] as an assimilation of the acoustic property of gravity from the vowel to the preceding consonant. The consonant spectra and the following vowel spectra share a predominant energy peak in the same frequency region and the spectrum of the burst for the velar and the palatal in the environment of [i] share a compact spectrum, although the frequency of the peak is higher in the case of the palatal. An assimilation is not found from [cu] to [ku], however, presumably due to the lack of shared acoustic properties between contiguous sounds as well as between the modified and original sounds. No account for the lack of assimilation from [ci] to [ki] is given, however.

In a conflicting cue task, Dorman & Loizou (1996) found that listeners often mistook [bi] for [di] when the [bi] stimuli had a relative spectral change appropriate for [di]. They attributed this result to the similarity of the formant transitions of [bi] and [di]. The reverse confusion did not occur, however. Fischer-Jørgensen (1954), in a tape-

splicing experiment involving Danish stops, also found that the presence of a [d] burst preceding the voiced formants of a [bi] syllable caused the report of a 'di'. The presence of a [b] burst preceding the voiced formants of a [di] syllable, however, did not cause a significant confusion toward 'b'. In a similar study of English stops [p, t, k], Schatz (1954) found that for a [k] burst combined with vowels other than the one before which it was pronounced, [k] was not the preferred judgment by subjects. This and other conflicting cue studies suggest that asymmetric confusions rely at least in part on ambiguity of the signal caused by the similarity of cues in certain environments. But, the fact that two segments are *similar* along various acoustic cues is not a sufficient explanation for the asymmetry of the resulting confusions.

Repp & Lin (1989) proposed that certain vowels may have *affinities* with certain consonant places by virtue of their articulatory and spectral properties. This explanation also relies on the notion of acoustic similarity, but it goes one step further in addressing asymmetry in confusions. They propose that the rounded back vowels /u/ and /o/ are most similar to labial stops, while high front vowels such as /i/ and /e/ are most similar to alveolar stops. Vowel affinities for consonant place may account for many of the asymmetries noted above, but not for the more symmetric confusion between [k] and [p] in the context of [u]. Additionally, the exact perceptual mechanism involved in this process remains unclear.

Others have presented similar explanations that depend on whether or not the vowel is known by listeners, but the results are sometimes contradictory. In a study of French stops, Bonneau et al. (1996) found that for her French listeners, knowledge of the vowel [i] had the opposite effect for dental and velar stops. Under both context

conditions, [k] and [t] were often confused for each another, presumably due to their spectral resemblance. When the vowel was specified, listeners were more likely to confuse [t] for [k]. When the vowel was unspecified, [k] confusions increased. These results tended to suggest that there were response biases in favor of dentals when the context was not known. Repp & Lin (1989) did not find this pattern in English, however. They found that knowledge of the vowels increased identifiability for [k] and [p] when followed by [u], and especially [k] when followed by [a].

Lindblom (1986, 1990) suggests that the interaction between auditory distinctiveness and the extremeness of articulation generates many patterns described as *markedness* relations between different segments. A language must find a balance between a phoneme's perceptual salience and the extremeness of the movement of the articulators involved in its production (see Section 6.3 for more explanation). Generally, Lindblom claims, extremeness of movement outweighs the perceptual salience of gesture. Thus, though [gi] may be more salient than [gu], [gu] is favored for its lower physiological cost. Speakers presumably learn these patterns in production and apply them to perception.

Another obvious explanation for asymmetries in consonant confusion is the relative frequency of the phoneme in the language (Lyublinskaya 1966). Vitevitch & Luce (1999) found that in word recognition tasks, non-words with frequent phonotactic patterns in the language are likely to cause more rapid and more accurate identification of the non-word heard. This result suggests that CV confusions may be less likely to occur when the CV segment is more frequent in the lexicon.

In contrast to explanations that focus on the causes of listener errors, Vitevitch & Luce (1999) claim that the frequency of a phonotactic pattern *facilitates* identification of that pattern. Therefore, this account still requires ambiguity of the acoustic signal to trigger confusions. In the current study, we are interested in particular when ambiguity in the signal is substantial enough to cause confusions in place of articulation, and the variations of ambiguity by CV context.

Although markedness, due to either phonological considerations or production constraints, may play a role in asymmetric stop consonant confusions, acoustic causes of confusions are shown to be primary in the case of [ki] → [ti] confusions by Chang et al. (2001). Building on arguments from Ohala (1983, 1985), the authors propose that the high front vowel neutralizes formant onset cues to stop place, leaving only the relatively non-robust spectral cues of the bursts. Velar and alveolar stops have similar burst spectra in the context of high front vowels, with the exception that velars have an *extra* mid-frequency peak. The mid-frequency cue is often degraded due to entropy, and in the case of [ki], likely to cause confusions for alveolar stop place. Other asymmetric confusions may also be triggered by an extra feature, as in the case of [ki] → [ti], but each asymmetric confusion must be investigated case by case.

6 Historical sound change in the laboratory

We have previously summarized important trends in stop place confusions in laboratory confusion studies. Many of the same CV pairs found to be confusable in laboratory studies are subject to parallel shifts in diachrony. In this section, we summarize a few of the cross-linguistic historical sound changes that mirror the stop place confusions

previously discussed. This section also presents a summary of two theories of sound change: Ohala's model of sound change (Ohala 1981) and the H&H model of sound change (Lindblom 1986), both of which support the notion that laboratory confusions can simulate phonetically motivated historical sound change. Both of these models also fundamentally assume that listeners perceive speech from the acoustic information directly, in contrast to models of speech perception that assume an intermediate articulatory representation, such as *analysis-by-synthesis* (Stevens 1960), the motor theory of speech perception (Liberman & Mattingly 1985), and the direct realist approach to speech perception (Fowler 1994).

6.1 Parallel historical shifts in stop consonant place

Shifts in consonant place of articulation are relatively rare historical sound changes across the world's languages. Nevertheless, many shifts in stop place have been observed that exhibit the same trends discussed in confusion studies across the world's language families. For example, alveolars are rarely subject to historical shifts in place, vocalic environment appears to trigger certain stop place shifts, and historical sound change is often asymmetric, corresponding to the asymmetric confusions in the laboratory. In this section, we present a small typology of shifts in stop place that mirror the confusions discussed in Section 1.4. In many cases, the shift in voiceless stop place is part of a larger shift across other manners of production, but we will focus here on voiceless unaspirated stops in CV context.

Velar/alveolar shifts

The most common historical shift between velar and alveolar place in voiceless stop consonants is a shift from a velar stop consonant to an alveolar stop, or more commonly, an alveolo-palatal stop or affricate, triggered by a high front vowel, palatal glide, or palatalization of the velar stop. This phenomenon, often called *velar palatalization*, is a common type of palatalization and a relatively common sound change in the world's languages (Guion 1998). The link between velar palatalization and the phonetic similarity between palatalized or fronted velars and alveolars is the subject of much discussion (cf. in particular Keating & Lahiri 1993; Guion 1996).

Language	Change	Environment
Slavic (1st palatalization)	k> tʃ g > ʒ x > ʃ	{j, ĭ, i, ɛ, ē}
Slavic (2nd palatalization)	k> tʃ d > dʒ x > s	{i, ɛ}
Indo-Iranian	k> tʃ g > dʒ $g^h > dʒ^h$	{i, e}
Bantu	k> ts g > dʒ	[j] + {i, e}
Old to Middle Chinese	k> tɕ $k^h > tɕ^h$ g > dʑ x > ɕ	[j] + {i, e}

Table 6. Examples of velar palatalization (from Guion 1998).

Stop place shifts in CV position in the opposite direction, from alveolar to velar place, is unattested, though alveolar stops have shifted to velars with some frequency in VC# position. Fricatives, on the other hand, are known to shift from velar to alveolar in CV position (cf. shifts from [x] in Spanish to [ʃ] in Ladino, Lathrop 1980). The asymmetry found in attested alveolar/velar historical sound changes mirrors the asymmetry of listener confusions in the laboratory between these two places of articulation in CV environments.

Bilabial/alveolar shifts

Cross-linguistically, bilabial stops in CV position often historically shift to alveolar stops when they precede a high front vowel (Andersen 1973; Chen 1973; Homburger 1949). In Sotho, bilabials undergo a synchronic shift to alveolar stops when they are labialized (Guthrie 1967-1970), presumably due to a previously present palatal element (Ohala 1978). Shifts from bilabial to alveolar stop place in CV position is usually more common for nasals, but occurs for voiceless unaspirated stops as well. Below are a few cases of bilabials shifting to alveolar place historically or synchronically in the context of [i], [j], or [w] (Table 7). In some cases, intermediate stages, such as [ptʃ] or [pl] may occur, and indeed the target shift from bilabial to alveolar may itself be intermediate in the chain from labial to palato-alvoelar (pj> ti> ts> tʃ).

Language	Change	Environment
Proto-Slavic to Teták	p > t b > d m > n	{ʲ, i, ɛ}
Tai to T'ien-chow	p >pj > tʃ	{l}>{j}
Latin to Italian	p > pʲ > tʃ	{l}>{j}
Gwari to Ganagana, Nupe	p > ts b > dz	{j}
Proto-Bantu to Xhosa, Zulu	p > tʃ	{i}
Classical Greek	p> t m> n	{jo}
Sotho	p > tʃ	{i, j}

Table 7. Examples of bilabial to alveolar shifts in stop place (data from Ohala 1978: Andersen 1973; Li 1977; Malkiel 1963; Guthrie 1967-70; Meillet & Vendryes 1924).

In many of the confusion studies summarized in Section 1.4, bilabial stops were misheard as alveolar stops when they preceded the high front vowel [i]. This is mirrored in the historical changes shown in Table 7 above. A phonetic explanation for the prevalent shift in stop place from bilabial to alveolar in the environment of a following high front vowel in confusion studies and perhaps in historical sound change is offered by Guion (1996), who proposes a direct shift from [pʲ] to [tʃ]. She suggests that the voicelessness of the bilabial triggers the perception of the frication of an palato-alveolar affricate.

Historical shifts of alveolar stop consonants to bilabial place are less common than shifts from bilabial to alveolar place, and are triggered by a different context. They occur cross-linguistically when a labial element, usually a [v], immediately follows the alveolar stop. No historical shifts from alveolar to bilabial place appear to be triggered

by a CV environment, though alveolar stops are known to shift to bilabial stops when they precede [w], as in the Latin prefix [dwi-] > [bi]. A parallel could be drawn instead to alveolars preceding rounded vowels, such as [u], in laboratory confusion studies. In confusion studies reviewed so far, however, [tu] is rarely confused for [pu].

Language	Change	Environment
Middle Indic to Gujarati	t > p d > b	{v}

Table 8. Example of alveolar to bilabial shifts in stop place (data from Grierson 1969:1906).

Historical shifts from bilabial to alveolar place are commonly triggered by a following high front vowel. Stop consonant shifts from alveolar to bilabial position in the same context are not attested. This historical trend is mirrored in the previously discussed stop confusion studies. In place identification tasks, listeners often mistook bilabial stops for alveolars when they preceded the vowel [i], but alveolars were rarely mistaken for bilabial stops in this same context.

Velar/bilabial shifts

Historical shifts from velar to bilabial place or from bilabial to velar place are common in the world's languages. They are common in both directions and are usually triggered by labialization or palatalization. Due in part to these facts, historical stop shifts between labials and velars appear to be part of a larger phenomenon of labials and velars patterning as a class, presumably due to their similar acoustics.

Labials may shift historically to either a palatalized or labialized velar in all environments, including CV position. These shifts parallel confusions in the laboratory of bilabials that precede a rounded vowel, such as [u], which are most often mistaken for velars, as we have seen previously.

Language	Change	Environment
Latin	$p > k^w$	#__, with k^w leading the second syllable
Proto-Siouan-Iroquoian to Seneca	$p > k^w$	All
Proto-Algonkian to Atsina, Yurok	$p > k$	All
Rumanian	$p > k^y$ $b > g^y$	All

Table 9. Examples of bilabial to velar shifts in stop place (data from Hock 1938; Chafe 1964; Haas 1969).

A bilabial to velar shift in place of articulation is also found in fricatives, for example, in Dutch, [f] shifted to [x] in stressed codas preceding [t], as in *after* to *achter* 'after' (Bonebrake 1979). Though there is no direct connection between this shift and the CV stop shifts discussed in this thesis, it does provide additional evidence that labials and velars often pattern as a class (cf. Ohala et al. 1978).

Historical stop shifts from labialized velars to a primary labial place of articulation are also attested across the world's languages (Table 10). The most common velar to bilabial stop shift is triggered by an alveolar stop that immediately follows. This particular sound change, however, is outside the scope of this study. I will simply refer the reader to the Latin to Albanian and Rumanian shift (Pagliuca 1982) in words such as Lat. *pectus* to Rum. *piept*, 'breast/ chest' or Lat. *lacte* to Rum. *lapte*, 'milk' and the

reverse shift attested in Venezuelan Spanish (Grierson 1969:1906) in words such as *pepsi* to *peksi*.

Language	Change	Environment
Proto-Indo-European to Greek	$k^w > p$ $g^w > b$	All
Latin to Rumanian	$k^w > p$ $g^w > b$	All
Proto-Mixe-Zoquean to Mixe and Tapchultec	$k^w > p$	All
Proto-Indo-European to Proto-Germanic	$k^w > p$	All

Table 10. Examples of velar to bilabial shifts in stop place (data from Ohala 1978: Meillet & Vendryes 1924; Longacre 1967; Bennett 1969).

Velar fricatives are also known to shift to labial fricatives in coda position when they occur in the environment of back, preferably rounded vowels, presumably due to the reinterpretation of the devoiced [w]-like transition between the back vowel and the following velar fricative (Nieuwint 1981). In modern day English orthography, the original velar fricatives can still be found in words such as *laugh*, *rough*, and *cough*.

Clearly, the historical shifts in place for stop consonants in CV position mirror many of the common listener misperceptions in confusion studies summarized in Section 5.1. In the following section, I summarize two theories of sound change that explicitly propose how a historical sound change may arise from the same misperceptions on the part of the listener that occur consistently in controlled laboratory studies.

6.2 Ohala's theory of sound change

Ohala's theory of sound change (1971, 1974, 1981, 1990, 1993) is perhaps best characterized by its emphasis on the importance of the acoustic auditory signal in speech perception. The acoustic signal is the only physical manifestation of communication between the listener and the speaker. The speaker must rely on the information in the signal to determine phonemic identity, though non-phonetic information, such as lexical frequency, gender, and sociological characteristics of the speaker, are also known to play a role. The importance of the acoustic-auditory information for the listener is particularly supported by the growing body of evidence showing that distant or physically unrelated articulatory events work together consistently, presumably to enhance the acoustic-auditory signal (Kingston 1992) (cf. Riordan 1977), who shows that lip-rounding and larynx height work together to lower F2, for example, creating a *grave* sound).

There are many ways in which the mechanical constraints of the speech production apparatus can leave its imprint on the speech signal. The listener, when decoding speech, may not be able to figure out which features of the speech signal are intended and which are unintentional distortions, especially in a situation in which very little top-down information is available, such as when the speaker is spelling out unfamiliar names over the telephone. In Ohala's model of sound change, listeners learn to expect synchronic covariation of phonetic events and must reconstruct the pronunciation norm intended by the speaker.

Sound change occurs when a listener fails to associate the separate co-occurring phonetic events and instead takes one or more of them as independent. This is an error of dissociation, or *hypo-correction*. If a listener mistakenly parses together two phonetic

events that should have been parsed separately, this is a false association or *hyper-correction*. In this model the speaker is responsible for the variations in production, but listeners, through perceptual parsing errors, are responsible for creating pronunciation norms that differ from those of the original speaker. Most of these errors are quickly corrected by the listener, who has access to other utterances and non-phonetic sources of information. Errors that do not get corrected, however, may become a characteristic of an individual speaker. Given the right set of sociolinguistic circumstances, this particular speaker may then transmit the new pronunciation to other speakers. If the transmission is extensive enough, it could very well become a linguistic sound change. The conditions must be just right for this to occur, as they presumably were in the documented cases of sound change mentioned above.

Stop confusions, for example, consistently reoccur in laboratory studies, such as those summarized above (Section 5.1) and are thought to arise from the inherent ambiguity of these sounds in the auditory domain. Ohala's model predicts that mini-sound changes, or listener perceptual parsing errors, are the seeds of major historical sound changes, and have the benefit of being elicited in a controlled setting. As Ohala (1993) explains:

> ...the ultimate check on any hypothesis about the cause of a particular sound change is to test the hypothesis in the laboratory. If particular sound changes are posited to have a phonetic basis then one should be able to duplicate the conditions under which they occurred historically and find experimental subjects producing 'mini' sound changes that parallel them. It is because of the posited phonetic character of sound change that a laboratory study is possible: were the initiation caused by grammatical and cultural factors, this would be more difficult or perhaps impossible (Ohala 1993: 261).

Ohala also notes that common sound changes attested independently in substantially the same form in unrelated languages, such as those summarized so far, are likely to arise from language-universal factors such as the physics and physiology of the vocal tract and the nature of the human perceptual system. Thus, the historical stop consonant shifts are likely to be due to the same acoustic and perceptual factors as the confusion studies that report parallel trends in stop place. If the factors for vowel-triggered stop place confusions are found in the current study, they will have implications for corresponding historical sound changes.

6.3 The H&H theory of sound change

Lindblom (1986, 1990, 1995) also believes that the acoustic signal does not directly encode articulatory or acoustic auditory invariance, but rather plays the role of supplementing the multimodal information already in place in the listener's speech processing system. His model of sound change, however, puts more emphasis on the role of the speaker.

For Lindblom (1990, 1995), sound change is analyzed in terms of a two-step process of variation and selection. Phonetic variations are said to arise from the ability of speakers to adaptively tune their performance to the various social and communicative needs that they associate with specific speaking situations. According to this model, sound change is adaptive: If forms arise that match the current values of the evaluation criteria better than the old forms, they are more likely to be phonologized. This account is very similar to the Ohala mini-sound change scenario. It differs in that misperception

by the listener is de-emphasized as the sole seed of sound change. Lindblom argues that misperceptions cannot be the sole triggers of sound change since there are several ways in which listeners gain access to the unnormalized representation of segments (the phonetics of the message), not simply via a perceptual error.

Lindblom shows that (1) the listener and the speaking situation make significant contributions to defining the speaker's task and (2) the task shows both long- and short-term variations. In the ideal case, the speaker will allow herself only as much coarticulation as the listener will tolerate. The speaker knows that for successful recognition, the signal must possess sufficient discriminatory power. Lindblom believes that the signal should be discriminatory to succeed at its task. This view is motivated in part by evidence that the human perception system is designed to cope with partial information. In Lindblom's (1990) own words:

> The (ideal) speaker makes a running estimate of the listener's need for explicit signal information on a moment-to-moment basis. He then adapts the production of the utterance elements (words, syllables or phonemes) to those needs. This occurs along a continuum with more forcefully articulated *hyper* forms at one end and less energetic *hypo* forms at the other. As the performance level increases from hypo to hyper, both the duration and the amplitude of articulatory gestures tend to increase, whereas their temporal overlap tends to decrease. As a result, the context dependence of articulatory and acoustic patterns is minimal in hyperspeech and maximal in hypospeech. As further consequences, coarticulation and reduction are typical of the hypomode. And the vowels and consonants of hyperspeech are expected to be closer to their target values in hyperspeech (Lindblom 1990:1687).

Lindblom links his *hypo* versus *hyper* forms of online phonetic properties of speech to *weakening* (assimilations, vowel reductions, consonant deletions, and lenitions) and *strengthening* (vowel and consonant shifts toward polarization) in phonological processes.

In the task of the speakers in this study, since there is no explicit listener and the task is unnatural, hyperspeech is expected. Presumably then, these tokens will not show variation along this continuum, but they will show token-to-token variation within a small range of this continuum. Also, there is not much top-down knowledge available to the listener in the perceptual portion. One of the goals of the current study is to examine whether ambiguity and discriminability can be measured in the signal and whether confusions can be estimated from them.

7 The current research

It is the purpose of the research presented in this thesis to investigate the role of acoustic cues in listener confusions of stop place in voiceless, unaspirated, prevocalic stop consonants. This study differs in methodology from classical perceptual studies, in which generally one or two acoustic properties in synthetic or natural speech known to cue stop place are carefully manipulated. Instead, this study employs a large set of semi-automatically extracted acoustic features. Though some accuracy and faithfulness to human auditory processing is sacrificed with this method, it allows the relative ranking in discriminability by place of a large set of cues by CV context. Additionally, the research presented in this thesis uses only unaltered near-natural CV tokens in all acoustic analyses and perception experiments to study the natural variation in CV tokens. One

problem that this thesis hopes to avoid is the conflicting results of previous studies that may be partially due to the use of synthetic stimuli. Also, confusions induced by naturally occurring CV tokens strengthen claims that these confusions are induced by the same mechanisms that lead to historical sound change.

The approach of the current thesis is as follows. First, the method of semi-automatic feature extraction is presented and the mean values for each CV context are reported by feature (Chapter 2). The faithfulness of the feature extraction to the intended cue to stop place is discussed in each case. In Chapter 3, the feature values form a set of test and training data from which decision trees, a machine learning algorithm, classify stop place. Decision tree classification accuracies provide an estimate of the relative role of features in stop place identification by CV context. In Chapter 4, the results are reported from a perception study in which the listeners were played CV tokens that are either *canonical* or *non-canonical* along relevant acoustic features. The results of the experiment showed that listener performance in stop place identification depends at least in part on the amount of discriminatory information in the signal. In addition, listeners were found to rely on certain features that were predicted to be useful by decision trees for stop place identification. Chapter 5 evaluates several explanations for asymmetric confusions, also using results from the perception study. Frequency of a given segment and asymmetries at the feature level were found to affect the rates and direction of listener errors in stop place identifications. Finally, certain implications for the results found in this thesis are discussed and ongoing work in this area is presented.

Chapter 2

Extracting acoustic cues to stop place

1 Introduction

Acoustic cues to stop place of voiceless unaspirated stops in #[s]CV position reside in the relative amplitude (Cooper et al. 1952, Ohde & Stevens 1983) and spectral characteristics of the burst and any fricative phase that may follow (Blumstein & Stevens 1979, Cooper et al. 1952, Stevens 1999), as well as in the trajectory and duration of formant transitions (Dorman 1996) and voice onset time before the following vowel (Klatt 1975). The current thesis investigates the relative ranking of these acoustic cues by CV context using a database of 913 unaltered CV tokens uttered by English speakers. This chapter presents the methodology for extracting the acoustic properties known to correspond to stop place as well as some results from the cue extraction by CV context.

The current study's acoustic analysis of English stop consonants begins with the description of a corpus of voiceless, unaspirated English stops. These tokens were extracted from #[s]CV position and then analyzed for the values of the acoustic features found in previous perceptual studies to be relevant in stop place detection, including: (1) gross shape of the burst spectra, (2) relative energy and power of the burst with respect to the following vowel, (3) presence or absence of multiple burst releases, (4) voice onset time (VOT), and (5) formant onsets and transitions. The results from the feature value extraction are presented feature by feature in Section 3.

2 Recording and labeling CVs

Over 1500 CV tokens (where C is [p], [t], or [k] and V is [i], [a], or [u]) were extracted from the careful speech of seven American English speakers (four men and three women), ages 21 to 28. The majority of the speakers were native Californians, except for one woman and one man who were native Texans. Each speaker was recorded in a sound-treated room uttering the frame sentence *Take ___, for example* (Table 1). The target words were [s]-initial words (Table 1) chosen to elicit voiceless unaspirated stops. The subjects were presented each word 10 times or more on randomized index cards. Subjects were paid for their participation. The recordings were conducted in two sessions for each speaker, three months apart. Careful speech was used because hyper-articulated speech is thought to provide more discriminatory cues than fast or casual speech (Sönmez et al. 2000). Although careful speech is the least likely to cause confusions when presented to subjects unaltered, it provides the best environment for extracting features thought to be stop place cues.

CV context	[i]	[a]	[u]
[p]	*speak*	*spot*	*spook*
[t]	*steep*	*stop*	*stoop*
[k]	*skeet*	*Scott*	*scoop*

Table 1. Subject word list. Each word was uttered in the target sentence Take ___, for example.

A drawback to using native Californian and Texan speakers is that their [u]'s are notoriously unrounded and centralized. However, listeners of the same dialect were available in greater numbers for the perception experiment portion of the study.

The recordings were digitized at a sampling rate of 16,000 bits per second. Bursts (location and number), voicing onset, and transitions into the following vowel were hand-marked on the waveform of each CV token by the author (Figure 1). The temporal labels were aligned to each target CV, allowing acoustic information to be automatically extracted from the digitized signal.

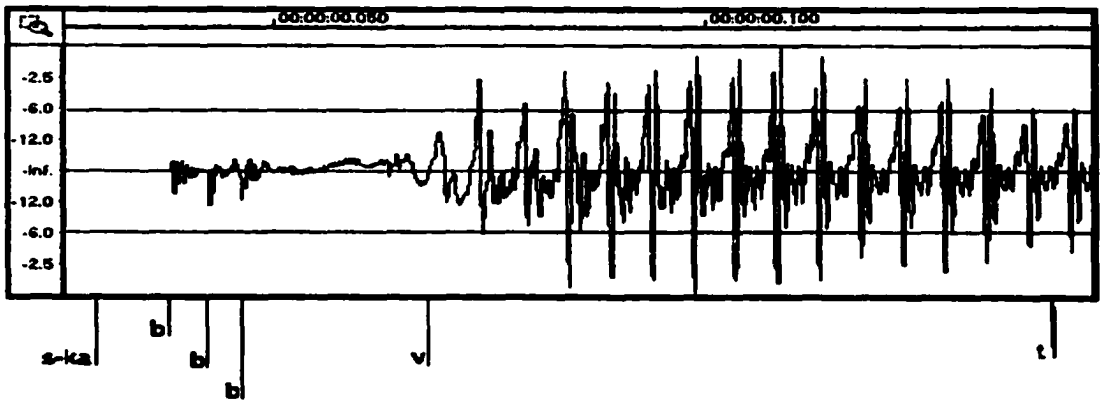


Figure 1. Hand-labeled waveform. This elicited [ka] token was hand-labeled for bursts, vowel onset, and final transitions.

The *start* of the target CV token was labeled with an 's' 10 to 30 msec before the consonant burst (but following the [s]), as well as a description of the consonant and vowel environment (e.g., 's-ka' in Figure 1). The burst (or bursts, in the case of stops with multiple bursts) for each unaspirated stop was labeled with a 'b', based on visual inspection of the waveform and the spectrogram. In the case of multiple bursts, all

calculations, including VOT and relative amplitude, were derived from the first burst only.

The onset of the voicing for the following vowel was marked with a 'v'. The onset of the vowel was consistently chosen as the first zero-crossing after the first mimimum or maximum of periodic structure (see Figure 1). The approximate beginning of the subsequent vowel transition into the following segment was marked by a 't' at the position where the vowel's formant structure began to change visibly, either by formant transitions into the following segment or by irregular glottal pulses (evidence of *creak*), which was common in [t]-final words.

For each CV token, information about the CV uttered, the speaker, the times and durations of the hand-labeled acoustic events, and the values of the specific acoustic features thought to cue stop place were automatically extracted from the temporal labels. The CV tokens were altered as little as possible for the perception portion of the experiment (see Chapter 4 for more details), since any change to the acoustic signal or environment (filter, noise, presentation level, etc.) may significantly alter confusion patterns as well as overall intelligibility scores (Bell et al. 1989).

3 Place cues of English stops [p], [t], [k]

The following sections examine the predicted acoustic features in turn, presenting both the methods of feature extraction from the signal and the mean values by CV for the following acoustic features and feature classes: (1) gross shape of the burst spectra (Section 3.1), (2) relative energy and power of the burst with respect to the following

vowel (Section 3.2), (3) presence or absence of multiple bursts (Section 3.3), (4) voice onset time (VOT) (Section 3.4), and (5) formant onsets and transitions (Section 3.5).

3.1 Burst spectra

Several investigators have noted that the spectral characteristics at the consonantal release provide cues to place of articulation (Cooper et al. 1952, Winitz et al.1972, Blumstein & Stevens 1978), especially when more reliable cues, such as formant transitions, are obscured. Indeed, many researchers concluded that the spectral properties of the burst hold cues to stop place that are *invariant* to the surrounding environment. In particular, Blumstein & Stevens (1979) found that for a spectrum calculated from an 8 msec window from 0 to 5 KHz, bilabials had a *diffuse-falling* spectrum (majority of energy in the low-frequencies), alveolars had a *diffuse-rising* spectrum (majority of energy in the high-frequencies), and that velars had a *compact* spectrum (a prominent spectral peak in the mid-frequency range from 1 to 3 KHz, depending heavily on the following vowel). In addition, the center frequency of the spectrum at the burst can hold cues to the place of articulation of the stop, since it corresponds roughly to the front-cavity resonance (Stevens 1999). In the current thesis, the gross shape of the spectrum at the burst is automatically estimated with a linear fit (for the spectral tilt) and a triangular fit (for the spectral peak) to the spectrum. These resulting spectral features can be combined to classify stop place of the CV tokens.

Extraction of the gross spectral shape of the burst

In an attempt to categorize each of the CV burst spectra according to the Stevens & Blumstein (1979) classification of diffuse-rising, diffuse-falling, and compact, an LPC spectrum of order 10 was generated from 0 to 8 KHz using a 6 msec Hamming window centered at the CV burst 'b', after the signal was pre-emphasized at 0.95. The spectrum was then sampled at 100 points along the frequency axis. A linear and triangular fit (two lines meeting at a node) were automatically generated for the sampled spectrum using a least-square fit from 0 to 5 KHz only, to replicate the Stevens & Blumstein (1979) study (Figure 2). The linear and triangular fits were designed to capture the overall tilt of the spectrum and the center frequency of the burst, or the most prominent peak. Derived features from the linear and triangular fits include the slopes, y-intercepts, and mean-squared error of the linear and triangular fits. In the case of the triangular fit, the optimal intersection for the two lines is called the *node*; its location in the frequency range is also included in the list of derived features. The differences in gross spectral shape by stop place (diffuse-falling for bilabial, diffuse-rising for alveolar, and compact for velar) are predicted to be captured by a combination of the features derived from the linear and triangular fits.

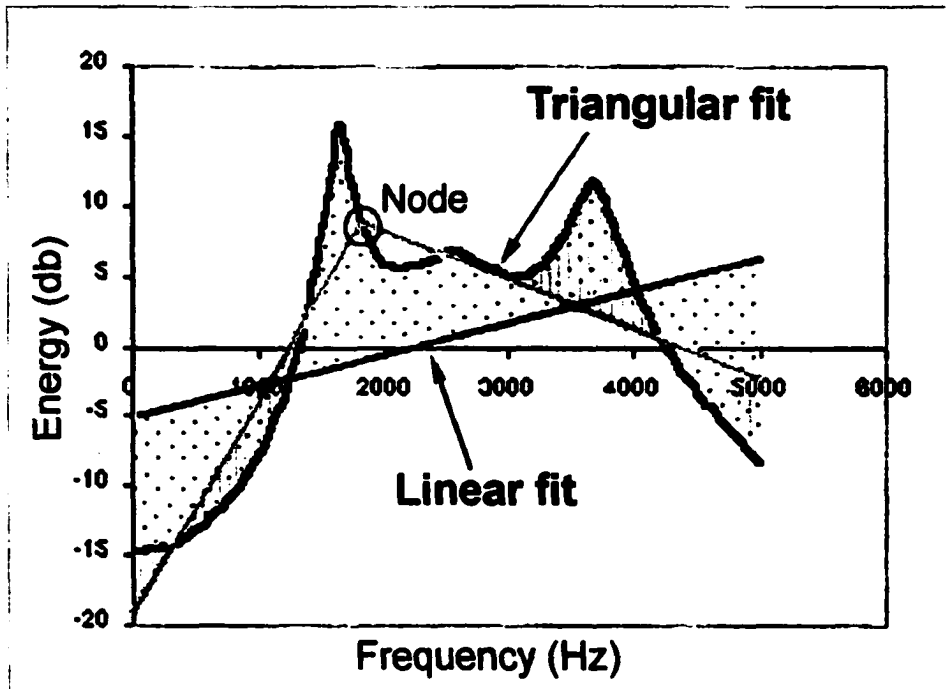


Figure 2. Sample burst spectrum and fitted lines. The average energy of the spectrum is centered at 0 db. The dotted area represents the error of the linear fit. The striped area represents the error of the triangular fit. The triangular fit is made up of two lines that meet at a node.

The gross spectral shape of bilabial stops, diffuse-falling, is predicted to be captured by a flat or falling linear slope. The pre- and post-node slopes for the triangular fit are also excpected to be relatively flat (low values, regardless of sign) and a triangular node in the low frequencies (Figure 3). In the examples shown here, the center frequencies of the bilabial burst spectra vary greatly by vocalic context. [pi] has the most prominent peak of the three vowel contexts. As exemplified here, in many cases the slope of the linear fit to the spectrum is actually a low but positive value, most often due to a dip in the low frequencies. For the bilabial burst spectra, as well as for other stop places, the triangular node fits slightly to one side of the center frequency of the burst. The node

feature is therefore not an accurate measure of center frequency, though it does indicate an approximate range of possible center frequencies for each stop place.

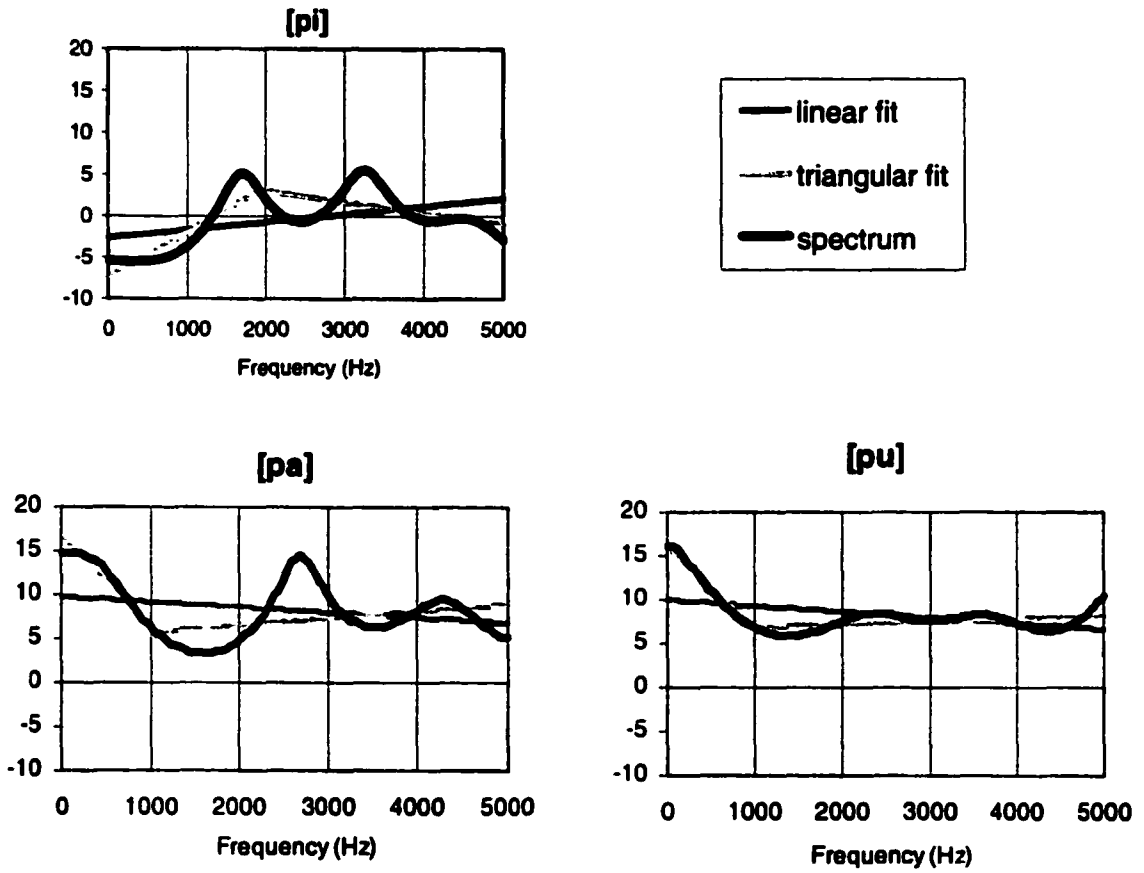


Figure 3. Burst spectra of sample bilabial bursts preceding [i, a, u]. Both the linear fit and the triangular fit to the spectrum are shown. All stops were produced by speaker RC.

The diffuse-rising shape attributed to alveolars is predicted to be captured by a steep linear slope and a triangular node in the high frequencies (Figure 4). The pre-node slope and the post-node slope are expected to be fairly steep, especially in the case of [ti] and [ta]. Note that in the case of [tu], the highest peak is immediately followed by a second peak, causing the steepest linear fit in these three examples. For alveolars

preceding [i], there is a prominent peak in the 3.5 to 4.5 KHz range. For [tu], this same range also contains an important peak, but usually with a greater bandwidth. The segment [ta] generally has the most energy in the highest frequencies, but that energy is distributed across three peaks.

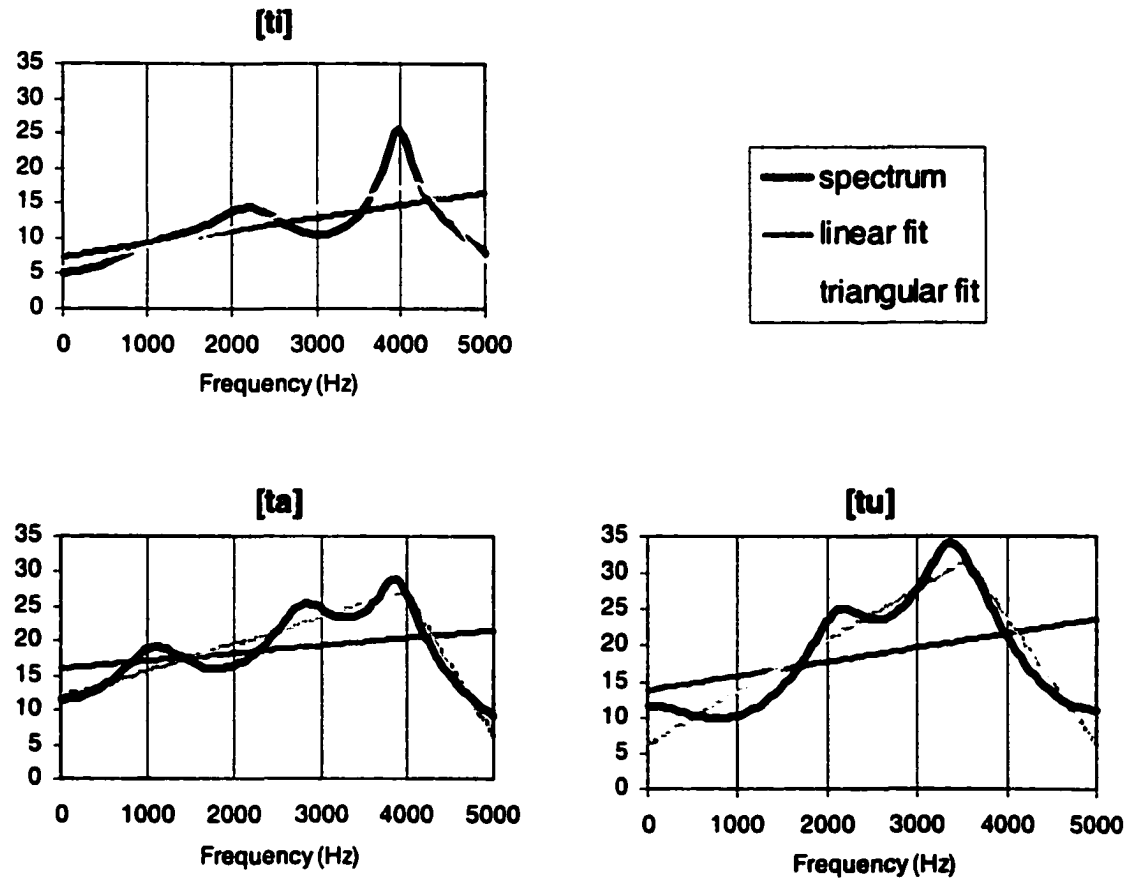


Figure 4. Burst spectra of sample alveolar bursts preceding [i, a, u]. Both the linear fit and the triangular fit to the spectrum are shown in each case. All stops were produced by speaker RC.

The slope of the linear fit to the spectrum is thought to play a lesser role in characterizing velars, since it varies greatly by vowel. Instead, velars are predicted to

have a triangular node in the mid-range frequencies, due to the compactness of the spectrum (Figure 5). Both the pre-node and post-node slopes are expected to be relatively steep, for the same reason. Note the relative prominence of the mid-frequency peak in the [ki] example.

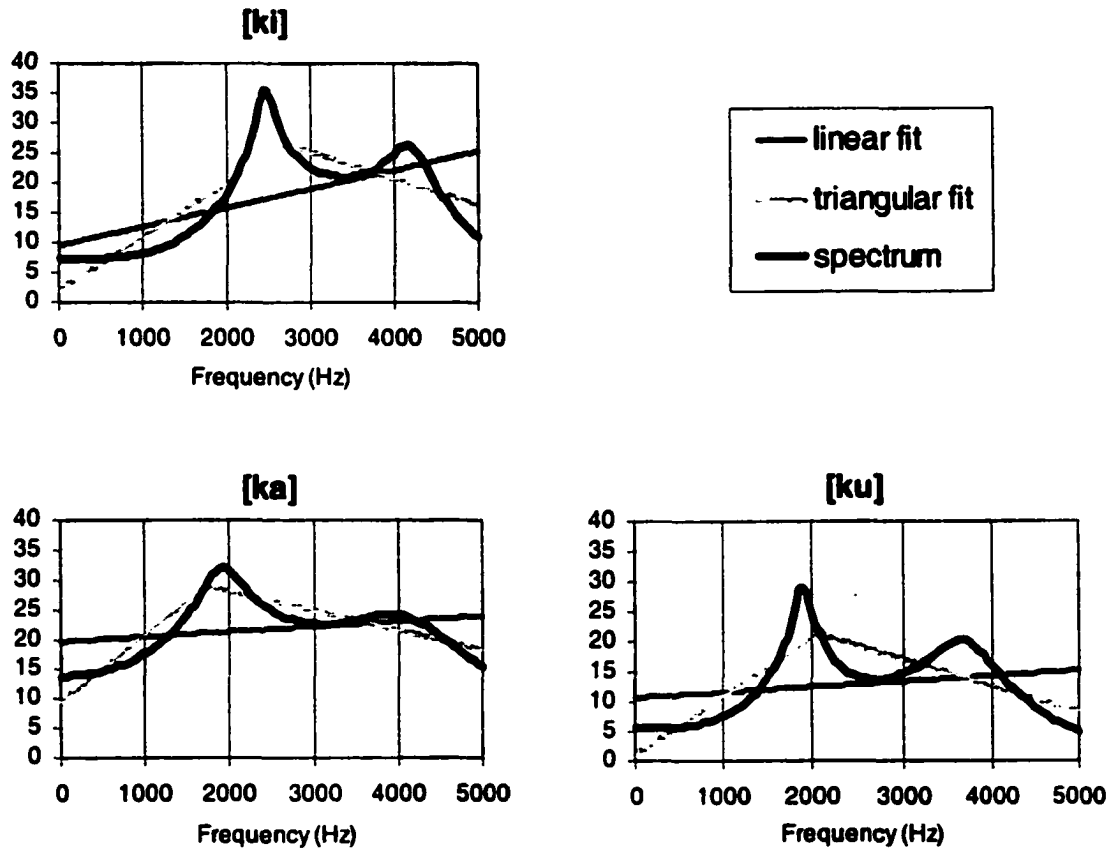


Figure 5. Burst spectra of sample velar bursts preceding [i, a, u]. Both the linear fit and the triangular fit to the spectrum are shown in each case. All stops were produced by speaker RC.

Gross spectral shape by CV context

The acoustic features derived from the spectral fits were extracted for each of the CV tokens collected. The average values for slopes and errors of both the linear and triangular fit as well as the average triangular node values are presented in turn by CV environment.

The slope of the linear fit to the burst spectrum is indeed flatter (lower in value) for bilabial stops than for alveolar and velar stops (Figure 6). Alveolars have the highest slopes, corresponding to their rising shape. Velar spectral slopes vary by vowel but are especially steep when the stop precedes [i], due to the large amount of high-frequency energy and the prominent peak in the mid-frequency range. In this context, velar stops are comparable to alveolar stops along this feature. Velar stops preceding back vowels have flatter slopes for linear fit, since their burst energy is distributed more evenly throughout the frequency range.

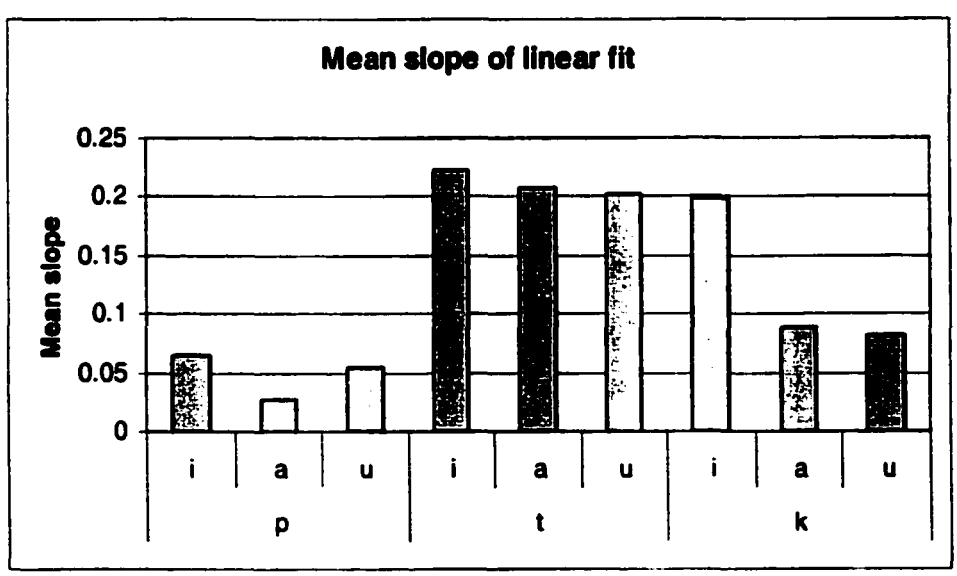


Figure 6. Mean slopes of the linear fit to the burst spectrum presented by CV context.

Further spectral information is provided by the values of the pre- and post-node slopes of the triangular fit (Figure 7). Both the pre- and post-node slopes are relatively flat for bilabials. As expected, the alveolars have the steepest (positive value) mean pre-node slopes of all the tokens. Again, the velars preceding [i] are more similar along these features to alveolars than to velars in other environments. With the exception of [ki], the velars showed flatter pre-node slopes than expected. The [ku] context yields an especially steep post-node slope because approximately a quarter of the cases have a node that fits to the secondary peak (since it is actually more prominent), causing a steeper post-node slope.

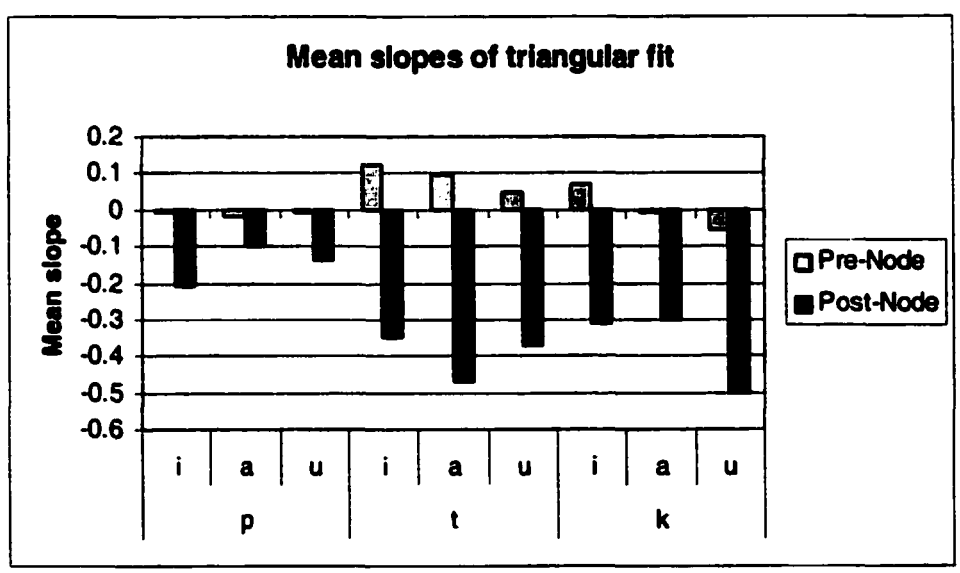


Figure 7. Mean slopes of the triangular fit to the burst spectrum by CV context.

The node of the triangular fit, our estimate of the center frequency of the burst spectrum, is predicted to be in the lower frequencies for bilabial stops and in the higher frequencies for alveolars. In the case of velar stops, the node is expected to vary by vowel. This prediction does hold true for this data despite the known inaccuracy of this

feature, which usually fits slightly to one side of the peak. For [pi], the node tends to fit either the first (~1700 Hz) or second (~3500 Hz) resonance, resulting in an average around 2600 Hz (Figure 8).

Velars in front of back vowels exhibit node positions in the lower frequencies on average. For [ka], there is a major spectral peak around 1800-2300 Hz. Most of the nodes of the triangular fits fit slightly to the left of this peak. For [ku], there are also two major spectral peaks, the first around 1500 Hz and the second ranging from 3.5 to 4.5 KHz. The mean for the node position is roughly the average of these two cases. In the case of [ki], the majority of the triangular fit nodes fit slightly to the right of the major spectral peak (2.5 to 4 KHz), yielding an average node value comparable to alveolars in the same context.

Alveolars, with the largest concentration of energy in the highest frequencies, have the highest node values on average. For [ti], the major prominent peak is around 3.5 to 4.5 KHz; the node fits to either side of this peak. For [ta], there are usually two spectral peaks with the node usually matching the high-frequency peak (around 3.2 to 4.5 KHz). In many cases, however, the node will match a low-frequency peak, yielding a slightly lower node average. For [tu], there is one major spectral peak around 3.8 to 4.0 KHz. The node fits to either the left or the right of this major peak but due to its slightly larger bandwidth, the mean is shifted upward in frequency. Velars show the most variation in node position by vowel context. In fact, [ki] is more similar to alveolars along this feature than to velars in other vocalic contexts.

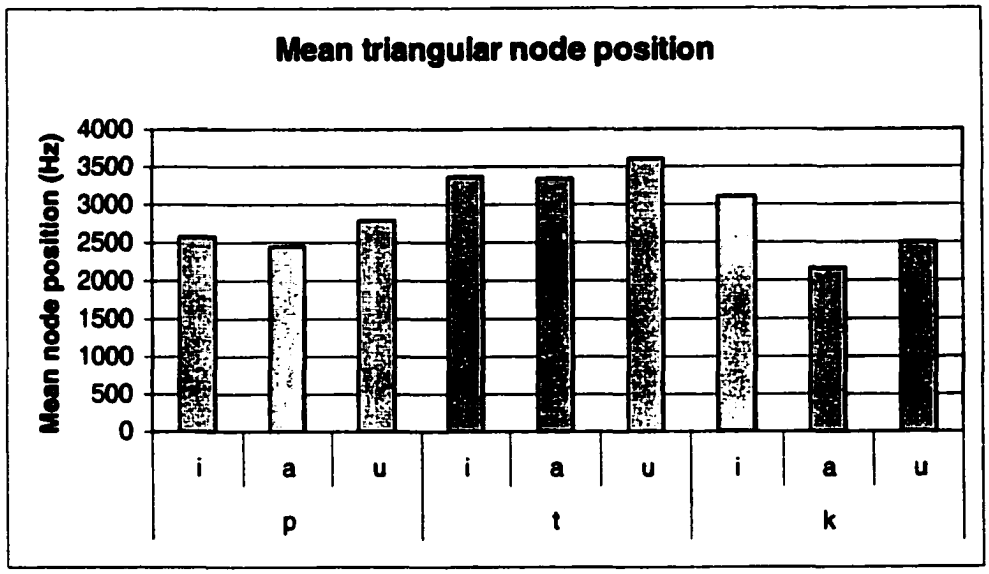


Figure 8. Mean node position of the triangular fit to the burst spectrum by CV context.

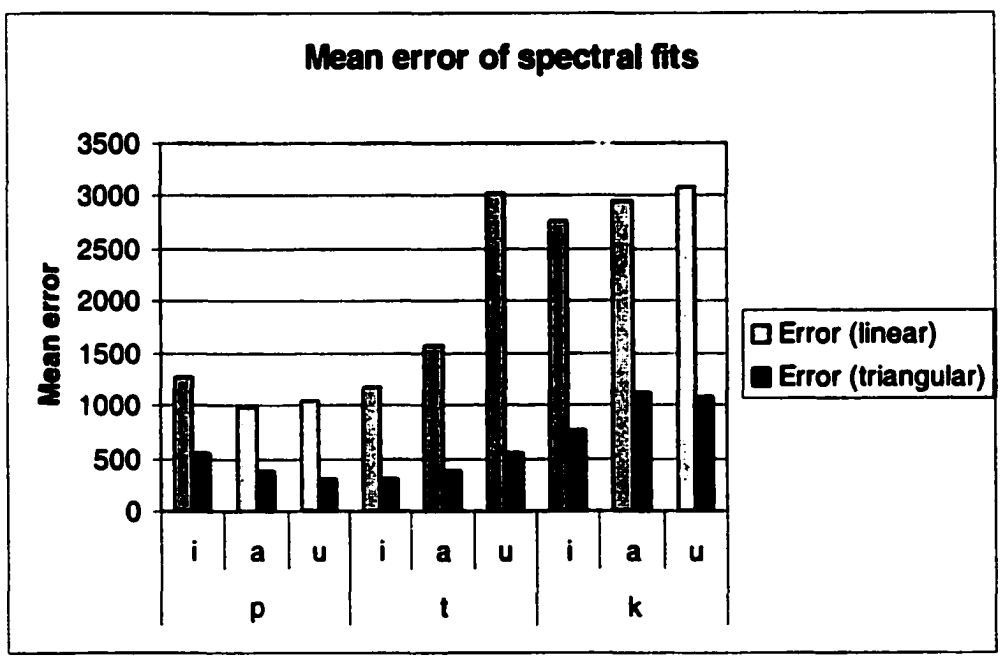


Figure 9. Mean error of the linear and triangular fit to the burst spectrum by CV context. The mean error is derived from the error of the linear and triangular fits to the burst spectrum after normalization by energy.

Triangular fits to both the alveolars and bilabial burst spectra yield low errors, as compared to velars (Figure 9). This indicates that linear error correctly distinguishes spectrally diffuse stop bursts from compact ones. The error of the triangular fit across all contexts is much lower than the linear fit error, but their trends are similar. The absolute difference between the linear and triangular fit error is greatest for velars and [tu].

With the exception of [tu], the error is relatively low for alveolar and bilabial stops. For [tu], the combination of a secondary peak in the lower frequencies and a prominent dip after the highest peak results in a very high error in both the linear and triangular fits to the spectrum. This was an unexpected result caused by the fitting method.

For [ta], there are two prominent spectral peaks, resulting in a high linear error. For [ti], there is a major prominent peak from 3.8 to 4.5 KHz. Because the fit is from 0 to 5 KHz only, the peak causes an overall rising linear spectrum, yielding a relatively low error. Both [ka] and [ku] have two major spectral peaks in the 0 to 5 KHz region, yielding high errors on average for the triangular fit. Because the second and third resonances are often merged in the case of [ki], a triangular fit aptly captures the gross spectral shape, yielding slightly lower errors for velars in this context than for [a] and [u].

A general problem with the linear error is that it fails to distinguish between a spectrum with a single prominent peak, as in the [tu] burst spectrum, from a spectrum with multiple, less prominent peaks, such as the burst spectrum of [ka]. The error of the triangular fit distinguishes between single and multiple peaks better, since it simply matches the main peak. It is still sensitive to peak prominence, however. This feature

could easily be improved in further studies by extracting the spectral maximum instead of a least-squares triangular fit.

A series of one-way ANOVA analyses were performed on each of the spectral features extracted from the spectrum at the stop burst to determine whether they varied significantly by consonant place, vowel, speaker, or gender of speaker. An additional univariate analysis was conducted to investigate variation by the interaction of both consonant place and vowel identity. All spectral features were found to vary significantly by consonant place (Table 2) and by the interaction of consonant place and following vowel.

Spectral Features		**Consonant Place** F(2,688)	**Vowel** F(2,688)	**Consonant * Vowel** F(4, 686)	**Speaker** F(6,684)	**Gender** F(1,689)
Linear Fit	**Slope**	207.69	17.9	13.05	8.99	19.64
	Error	131.50	11.49	18.96	14.56	31.58
Triang. Fit	**Node**	50.78	7.18	6.00	3.43	-
	Pre-node slope	74.92	24.11	11.35	25.84	94.20
	Post-node slope	34.77	-	5.59	6.04	-
	Error	151.67	-	16.77	-	-

Table 2. One-way ANOVA F values for spectral features. '-' indicates an F value that is not significant ($p < 0.01$). '*' indicates feature interaction. Higher F values indicate more highly correlated variation between the spectral feature and the factor.

Node and triangular slopes yielded the least significant variation by consonant place. The pre-node slope was particularly sensitive to vowel identity and gender of the speaker. Small but significant variation by speaker identity were found for all factors except the error of the triangular fit. Linear fit, linear error, and pre-node slope were also found to vary significantly by gender (Table 2).

Y-intercept values of the linear and triangular fits were not included in the final analysis since they were highly variable across individual tokens. This is presumably due to the fact that the signal was not normalized by amplitude. The y-intercepts were most sensitive to overall amplitude differences, which vary greatly by speaker and by token. As previously mentioned, the triangular and linear error exhibit similar trends.

No one spectral shape feature is sufficient for correctly classifying the three stop places from the burst characteristics. When used in combination, however, these derived features can distinguish between stop places, derived from their differences in the gross spectral shape at the stop burst.

Previous studies have suggested that humans use burst spectra characteristics for stop place categorization only when more reliable cues such as formant transitions are obscured or absent. The data from this analysis show that for the English CV tokens studied here, gross shape of the spectral burst can be automatically extracted from linear fits. The derived features (linear slope, node, etc.) vary predictably by stop place. This suggests that gross spectral shape is available to listeners as a cue to stop place.

3.2 Relative amplitude and power of the stop burst

The amplitude of the burst release and following frication, and in particular the high-frequency amplitude of a CV, has been observed to cue listeners to place of articulation in voiceless stops (Ohde & Stevens 1983, Repp 1984b, Plauché et al. 1997). Alveolars have the highest amplitude bursts, with most of the energy in the high frequencies. Bilabials are associated with a lower amplitude burst with respect to the following vowel. Velars tend to share the overall high amplitude of alveolars, but with less energy in the high frequencies.

In the current study, the relative amplitude and relative power of the bursts and following frication are calculated for each CV token. The distribution of that energy across the frequency domain is captured by the previous set of spectral features (especially node and slopes).

Relative amplitude and power extraction

To capture overall energy information, a Hamming window beginning 5 msec before the burst 'b' and continuing until the 'v' was used to calculate the sum of the absolute energy for the duration of the window. The result serves as a measure for the overall amount of energy during the burst and transient and any amount of aspiration until the onset of voicing. No pre-emphasis was performed on the signal for this feature. The overall energy from 5 msec before the burst to voicing onset was then normalized by the following vowel. The resulting normalized value, relative energy of the burst with respect to the vowel, can be compared across tokens and across speakers.

Normalizing with respect to the following vowel, however, presents a problem, since vowels vary in their inherent amplitude (Lehiste & Peterson 1959). The vowel amplitudes in the current study were therefore first normalized by the average amplitude across all speakers for a given vowel. The three vowels were adjusted by multiplying their summed energy from 20 to 30 msec after voicing onset by the following ratios: [i] = 1.42, [a] = 1.08, and [u] = 1.30, as derived from the Lehiste & Peterson (1959) study. The absolute amplitude of the burst and frication was then normalized by the adjusted amplitude of the following vowel, yielding the relative amplitude of the burst.

As we will see in Section 3.4, the length from 'b' to 'v' varies predictably by stop place. Thus, a second measure, relative power (energy per second) was calculated that would be sensitive not to the duration of this interval but instead to how much energy per second was generated on average by each stop release and following frication. Relative power was calculated by dividing the relative energy by the length of 'b' to 'v' plus 5 msec. The energy values were normalized by the length of 'b' to 'v', yielding a measure for absolute and relative power (energy per second) for each CV token.

Relative amplitude and power by CV context

As predicted by previous studies, the bilabial stop bursts have the lowest relative energy of all stop places shown here. The relative energy of bilabial bursts also shows less variation by vocalic context than both the alveolar and velar places. The velar stop bursts and alveolar stop bursts have larger amplitudes and more variation depending on vocalic context, with the alveolar stop bursts' amplitudes growing larger from [i] to [a] to [u] contexts and the velars varying in the opposite order.

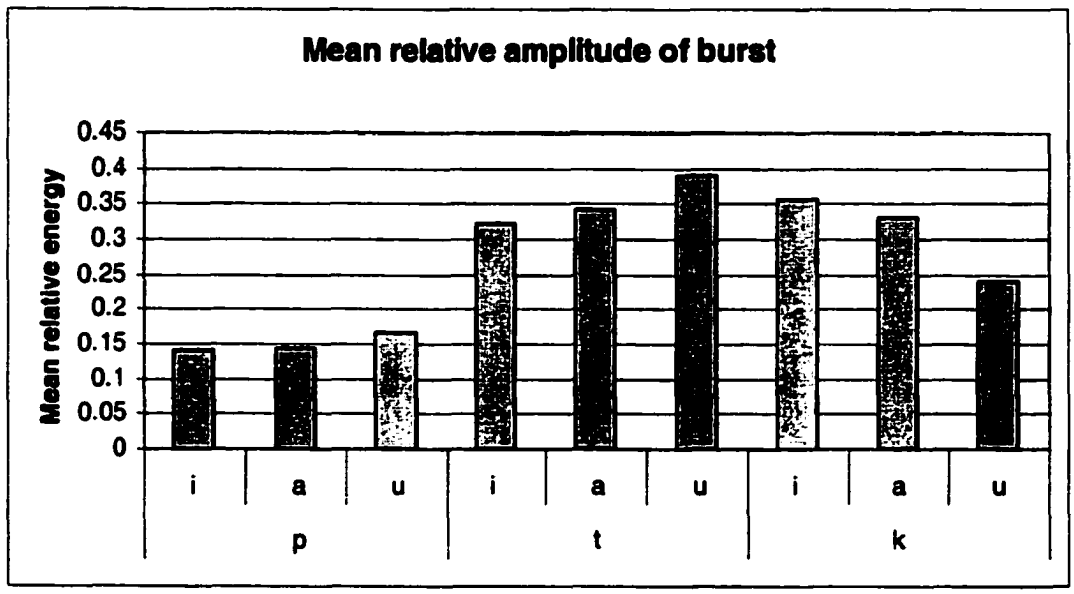


Figure 10. Mean relative amplitude of the burst and following frication with respect to the following vowel by CV context.

Trends seen in the mean relative energy by CV context are affected by variations of duration of stop release and following frication. Normalizing by release duration reduces the variations by stop place, as expected. Alveolar bursts have the most energy per second, especially when they precede back vowels. Velars in the context of /a/, however, show relative power levels that are comparable to those of alveolars. Velars show the most variation by vocalic context. Interestingly, the sum of the average relative energy of velars across the vowel contexts yields an average that is equivalent to the mean for bilabial stops. This may account for results in previous studies that reported results summed across vowel contexts where velars and bilabials had similar values for relative energy.

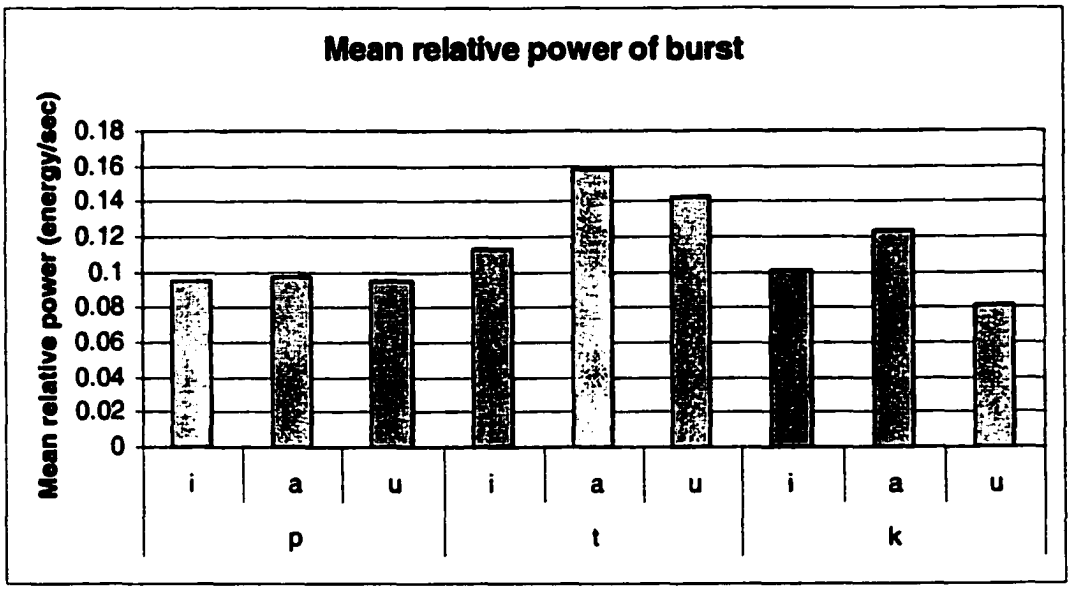


Figure 11. Mean relative power of the burst and following frication with respect to the following vowel by CV context.

A series of one-way ANOVA analyses were performed on the relative energy and the relative power of the stop release to determine any significant variation by consonant place, vowel, speaker, or speaker gender. The relative energy of the stop release was found to vary significantly by consonant place $F(2,688)=104.39$ and speaker $F(6,684)=9.63$, but not by vowel or speaker gender. The relative power of the stop release was found to vary significantly by consonant place $F(2,688)=29.79$, following vowel $F(2,688)=8.22$, and speaker $F(6,684)=11.56$, but not by speaker gender.

Due to the similarity of velars and alveolars along relative energy, this feature is predicted to be most useful for detecting bilabial place. Relative power, however, is predicted to be useful for listeners in detecting alveolar stops.

3.3 Number of stop bursts

Another acoustic feature of stop bursts that may be used by speakers to identify stop place is the occasional presence of multiple bursts. Such multiple bursts are most commonly associated with voiceless velar stops, presumably a result of the large area of the velar constriction and its relatively slow release (Stevens 1999). This phenomenon may be found with coronals and even bilabials before closely rounded vowels, but they are most frequent with articulations involving a large articulator and a large constriction area (velars and uvulars). No study to my knowledge has tested the role of number of burst releases in listener perception.

Multiple burst extraction

For each CV token, the number of bursts was hand-labeled (number of 'b' labels) by the author, relying on the visual output of both the waveform and the spectrogram. This feature is unique in its categorical nature; all other features use continuous values.

Multiple bursts by CV context

A histogram of the number of stop tokens from the database with 1, 2, 3, or 4 bursts (Figure 12) shows that though velars account for the majority of stops with multiple bursts, many alveolars have multiple bursts and a few cases of bilabials with 2 bursts exist. Additionally, not all velars in #[s]CV contexts are produced with multiple bursts. The ratio of velar stops with multiple bursts is especially low in spontaneous speech (Sömnez et al. 2000).

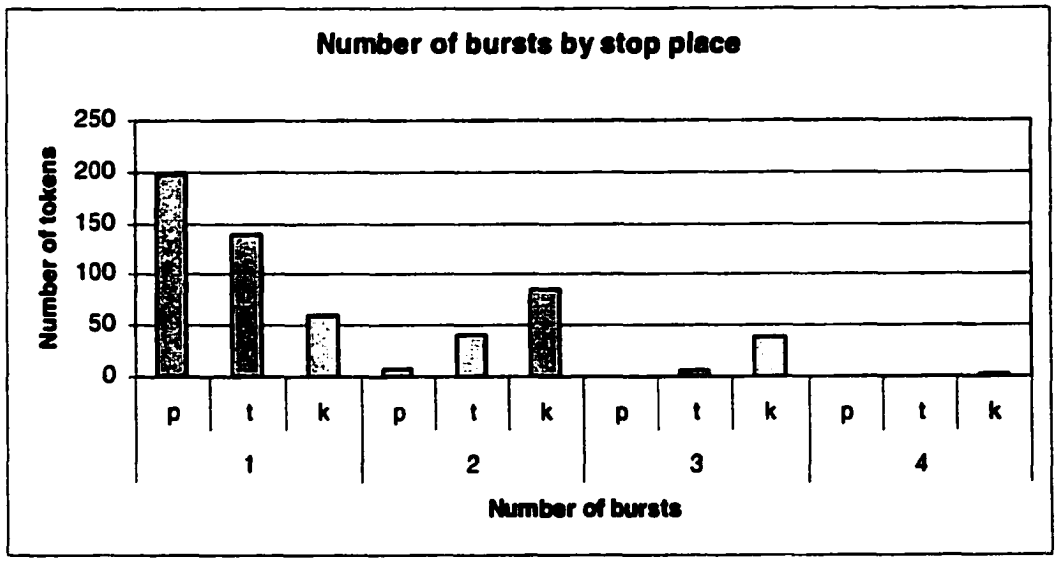


Figure 12. Histogram of the number of CVs with 1, 2, 3, and 4 stop bursts.

The frequency of multiple burst occurrences does not appear to depend on the following vowel, with the exception of velars, which show a slightly higher rate of multiple bursts in front of high front vowels (Figure 13). This would follow from Stevens' model of multiple bursts, which associates the probability of a multiple release with both a greater area of the constricted region and a slower rate of release. In the case of a velar stop and a following [i], the area of the constriction is expected to be slightly larger than for other constrictions, yielding an environment ripe for the Bernouilli effect (see Chapter 1, Figure 2).

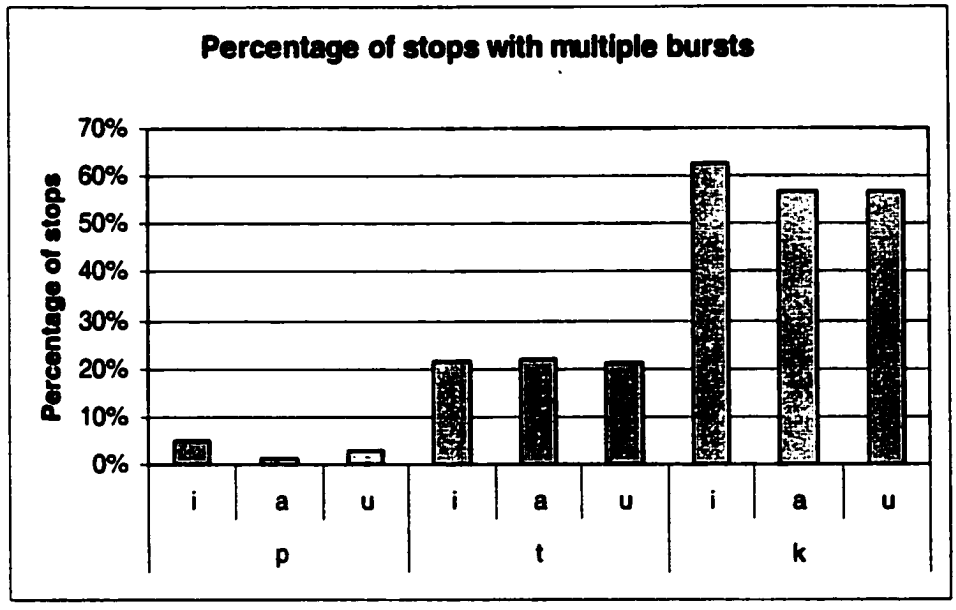


Figure 13. Percentage of CV stops with multiple bursts.

Results from a one-way ANOVA analysis showed that the number of bursts varied significantly by consonant place $F(2,688)=171.35$, but did not vary significantly by vowel identity, speaker, or speaker gender.

The categorical feature number of bursts may not directly cue listeners to stop place, since the auditory system is not thought to make temporal distinctions in such a short time domain (Delgutte 1980). It is hypothesized, however, that multiple bursts do enhance the spectral information of the burst by distributing it over a longer period of time. In the following chapters, we will estimate the relative importance of multiple bursts as a cue to stop place (Chapter 3) and test its relative ranking with a perception study (Chapter 4).

3.4 VOT

Voice onset time (VOT) is defined as the duration from the transient of the burst to the onset of periodic vocal fold vibration and the aspiration that follows (Klatt 1975). VOT is a primary cue cross-linguistically to the perception of stop manner, i.e., distinctions between voiced, voiceless, and aspirated stops. As we have previously discussed, however, VOT is also known to vary predictably with stop place from short to long as the constriction moves back in the oral cavity (p < t < k), making it a prime candidate for a cue to stop place.

VOT extraction

For the purposes of this study, the VOT value for each CV token was computed directly from the hand-labeled acoustic signal. As previously discussed (Section 2.1), voicing onset 'v' was defined as the first zero-crossing of periodic signal after a maximum or minimum (Figure 1). VOT for each CV token was computed directly from the subtraction of 'v' from 'b'; that is, the duration of time between the burst release and the voicing onset of the following vowel.

VOT by CV context

The mean VOT values by stop place closely follow trends found in previous VOT studies (Klatt 1975; Ohala 1981) (Figure 14). Voiceless bilabial stops have the shortest VOT, followed by alveolars, with velars exhibiting the longest VOTs.

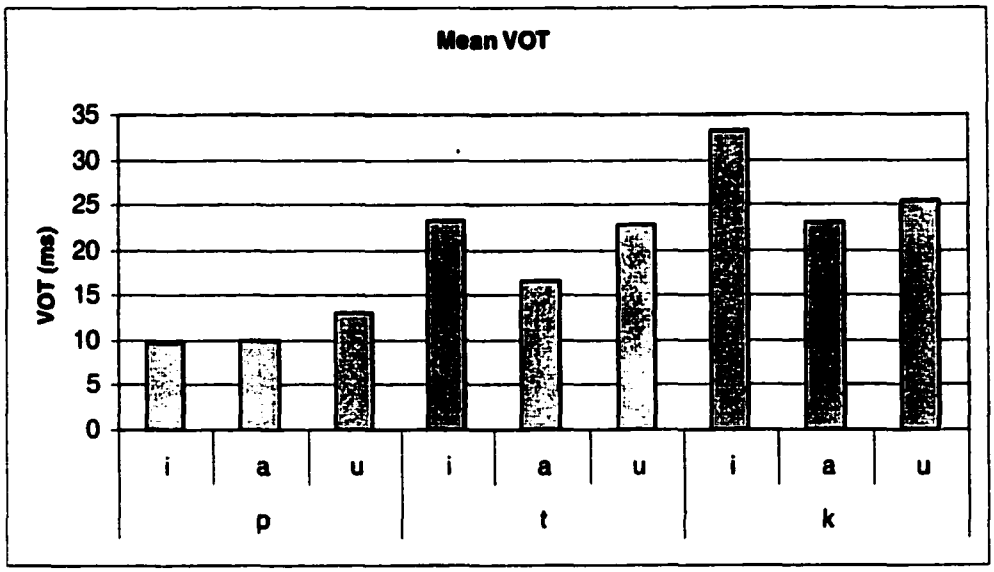


Figure 14. Mean VOT by CV context.

VOT is also found to vary with the height of the following vowel in the CVs studied here, as in previous VOT studies (Lehiste 1970; Klatt 1975; Chang 1999, but cf. Lisker & Abramson 1967). Both alveolar and velar stops have longer VOTs on average when they precede the high vowels [i] and [u] than when preceding [a]. Bilabial stops have a longer VOT on average preceding the vowel [u] than in other vocalic contexts.

A one-way ANOVA analysis was conducted to examine whether VOT was affected by consonant place, vowel identity, speaker identity, and gender. VOT varied significantly by consonant place $F(2,688)=280.75$, as predicted. VOT also varied significantly by the following vowel $F(2,688)=19.24$ and by speaker $F(6,688)=9.48$. VOT did not vary significantly by gender, however.

Though the role of VOT as a perceptual cue to stop place is thought to be secondary to features such as stop burst shape and energy, or formant transitions after the release of the stop, VOT varies predictably according to stop place and following vocalic

context. Therefore, VOT is predicted to be used by listeners as a cue for stop place. In the following chapters, we will estimate its relative importance as a cue to stop place (Chapter 3) and test its relative ranking with a perception study (Chapter 4).

3.5 Formant transitions

During the release of the consonant in the production of a CV sequence, the lips and tongue body begin to move toward the target configuration for the production of the following vowel. The movement of the articulators results in formant transitions that vary according to the place of articulation of the following vowel. The onset and rise time of formant transitions and, in particular, F2 transitions are considered primary cues for stop place (Cooper et al. 1952; Delattre et al. 1955; Liberman et al. 1967; Kewley-Port 1982, 1983).

Formant extraction

Formant values (F1, F2, F3, and F4) for each CV token were extracted from the onset of voicing of the vowel 'v' to 50 msec into the vowel at 10 msec intervals. The ESPS formant tracker used a cosine window of length 49 msec to track the four formants with an LPC estimate of order 12. The formant tracker advances every 10 msec to perform the formant estimation. The signal was first pre-emphasized at 0.95. To estimate the rise time for each of the formants, a line was fitted from the onset value to 20 msec into the vowel. The slope of the linear fit to the formant onset was automatically calculated for each CV token.

Formant transitions by CV context

In the current study, the method used by the automatic formant tracker proved to be more error-prone than hand-labeling. The automatic formant tracker used the values of the sample signal 10 msec before and 10 msec after each sampled point in order to extrapolate to the target value. This causes more errors at the onset of voicing when overall amplitude is low and very little formant in formation is available in the 10 msec preceding the voicing onset. Unfortunately, these problems occur precisely in the portion of the signal that is thought to hold the most information about stop place. Despite problems with the formant tracker at voicing onset, the automatically extracted formants do exhibit the predicted place distinctions.

Formants following bilabial stops are predicted to rise from voicing onset to the full vowel. The rise is caused by the opening of the lips, which when closed act as a formant depressor (Stevens 1999). Instead, as shown in Figure 15, for all mean automatic formant tracks, the bilabial stops show a steady onset or even a slight fall, with the exception of the F1 in [pa] and the F2 in [pi]. This is most likely due to the marker 'v' being slightly later than in other formant studies. Here, the onset of voicing was defined by the first zero-crossing after a minimum or maximum was achieved for the periodic signal (Figure 1). In other studies, the voicing onset is often the first zero-crossing or the first periodic maximum or minimum. This may account for the flat onset of the bilabial formants.

Note the fairly high mean F2 values in [pu], as well as in [tu] and [ku]. The [u]'s of the native Californian and Texan speakers are less rounded and more centralized than the cardinal vowel [u].

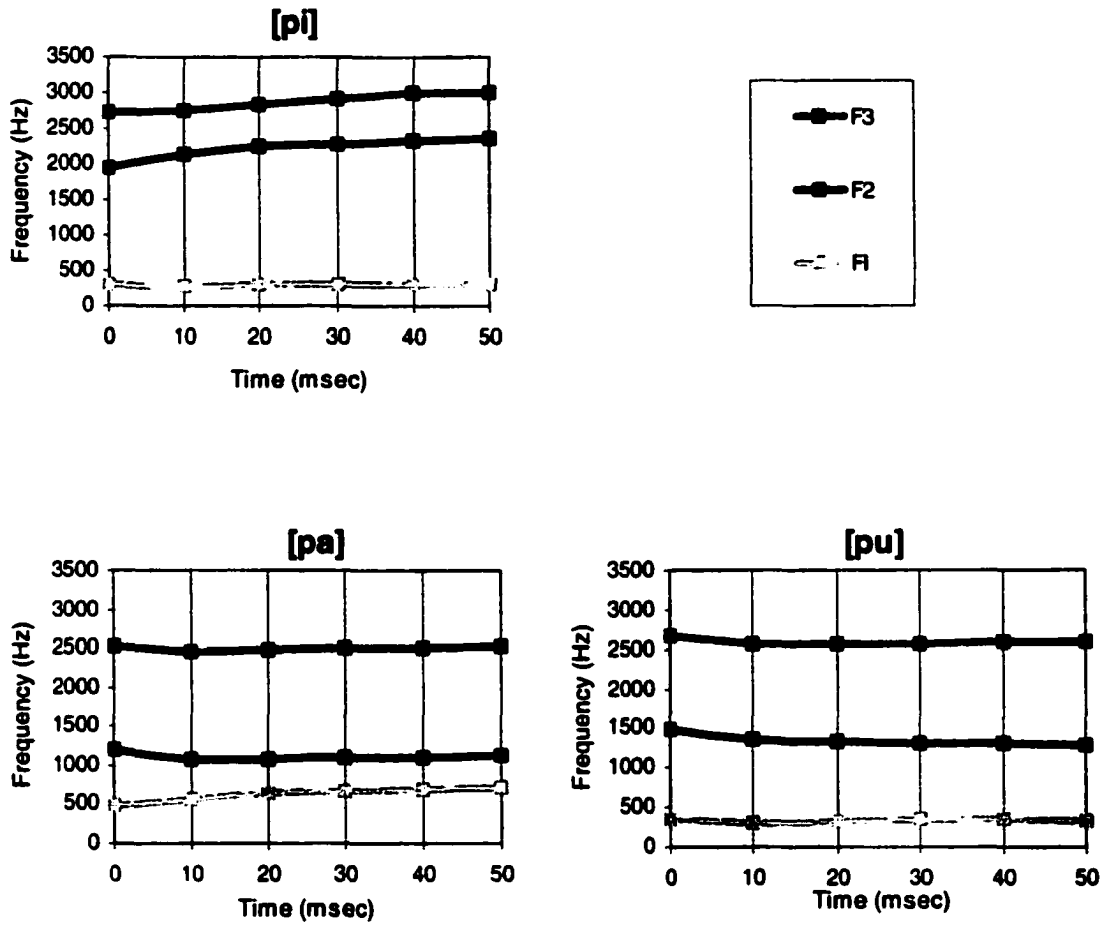


Figure 15. Mean F1, F2, and F3 values for bilabial stops automatically extrapolated at 10 msec intervals from 0 to 50 msec after voicing onset.

As predicted, formant transitions from alveolar stops all show a slight fall in F3 as they progress from voicing onset into the following vowel (Figure 16). The rise of F1 in [ta] is much slower than for [pa], which was mostly complete within 20 msec after

voicing onset. The formant transitions for both [ti] and [tu] are flat.

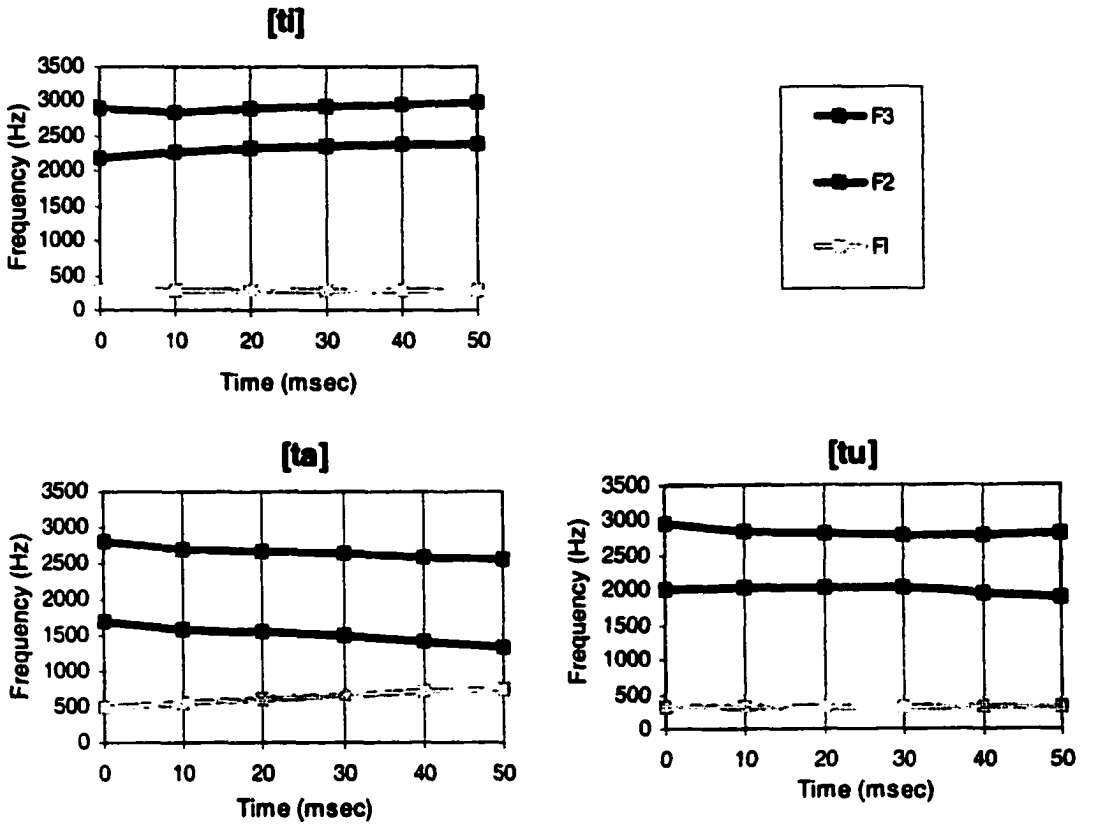


Figure 16. Mean F1, F2, and F3 values for alveolar stops automatically extrapolated at 10 msec intervals from 0 to 50 msec after voicing onset.

As in the alveolar CVs, mean formant transitions from the velar stops to a following [i] or [u] are fairly flat, differentiated only by the frequency range of F2 (Figure 17). The mean formant transitions of [ka] continue throughout the first 50 msec of the vowel; they are also the slowest formant transitions of the three stop places examined. With the exception of a slight rise in F2 in the case of [ti], the formant transitions for [ti] and [ki] are very similar. The formant transitions for [tu] and [ku] and [ta] and [ka] are

also similar in values and slopes. In CV tokens where V is [i] or [u], subjects are expected to rely on cues other than formant transitions to identify stop place.

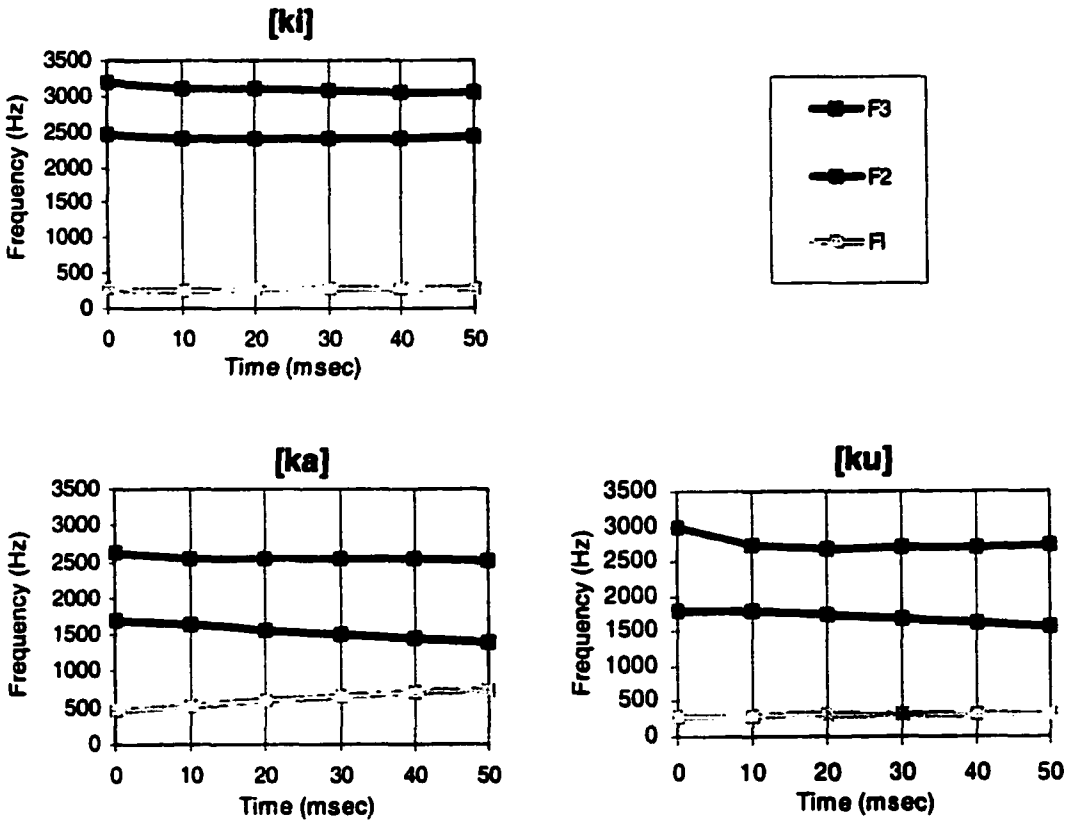


Figure 17. Mean F1, F2, and F3 values for velar stops automatically extrapolated at 10 msec intervals from 0 to 50 msec after voicing onset.

A series of one-way ANOVA analyses were conducted to determine whether the formant features extracted in this study varied significantly by consonant place, vowel, speaker, or gender of speaker. The results are shown in Table 3.

Formant features		Consonant Place F(2,688)	Vowel F(2,688)	Speaker F(6,684)	Gender F(1,689)
F1	'v'	-	126.66	9.52	-
	'v' + 10	-	1962.55	-	-
	'v' + 30	-	3527.64	-	-
	Slope at 'v'	-	19.76	6.27	-
F2	'v'	92.51	219.23	6.52	30.50
	'v' + 10	77.58	423.52	8.56	31.96
	'v' + 30	39.77	573.22	7.36	23.42
	Slope at 'v'	-	30.09	-	-
F3	'v'	39.04	34.88	6.25	29.90
	'v' + 10	28.09	65.97	8.76	27.56
	'v' + 30	8.65	111.35	9.76	37.03
	Slope at 'v'	-	9.66	-	-
F4	'v'	43.71	-	15.64	88.10
	'v' + 10	33.92	-	50.42	208.19
	'v' + 30	5.49	9.94	65.02	279.25
	Slope at 'v'	8.10	-	8.00	8.84

Table 3. One-way ANOVA F values for formant features. '-' indicates an F value that is not significant ($p < 0.01$). Higher F values indicate more highly correlated variation between the formant feature and the factor.

In general, the slopes yield little information about the consonant place or the identity of the following vowel. This is most likely due to a combination of the poor formant tracker estimates and the relatively late position of 'v' during the labeling of each CV token. The slopes also yield little information about the gender of the speaker. They were excluded from further analyses.

F1 contains information about the identity of the following vowel, but not about the consonant place, the speaker, or the gender of the speaker. The higher formants, especially F4, contain information about the speaker's gender and identity. F2 at the onset of voicing and 10 msec later varies the most significantly with consonant place, as predicted by other CV acoustic studies.

F2 is predicted to play a primary role in listeners' distinctions between bilabial and other stop places. Alveolar and velar stops, especially when preceding [i] and [u], are too similar along formant values for them to play an important role. In the following chapters we will investigate the relative role of the formant features examined in this section as cues to stop place.

4 Conclusion

This chapter has outlined the methods and results of the extraction of primary acoustic properties of stop consonants in CV position that listeners are known to use when detecting place of articulation. Specifically, various DSP techniques were used to semi-automatically extracted features from the acoustic signal of each of the CV tokens recorded and labeled.

Table 4 recapitulates the relevant acoustic features along with the specific CV contexts for which they are thought to be most useful and the abbreviation used for the feature throughout the following chapters. With a few exceptions, the mean values of the extracted features by CV vary predictably by stop place and often by the following vocalic context.

Acoustic features		CV Cued	Abbreviation
Linear fit to burst spectrum	**Slope**	Low – All [p]'s, [ka], [ku] High – All [t]'s, [ki]	LIN_SLPE
	Error	Low – All [p]'s, [ti], [ta] High – All [k]'s, [tu]	LIN_ERR
Triangular fit to burst spectrum	**Node**	Low – All [p]'s, [ka], [ku] High – All [t]'s, [ki]	NODE
	Pre-Node Slope	Low – All [p]'s, [ka], [ku] High – All [t]'s, [ki]	TRI_SLPE
	Error	Low – [pa], [pu], [ti], [ta] Med – [pi], [tu] High – All [k]'s	TRI_ERR
Relative amplitude of burst		Low – All [p]'s High – All [t]'s, all [k]s	REL_AMP
Relative power of burst		Med – All [p]'s, [ti], [ki], [ku] High – [ta], [tu], [ka]	REL_PWR
Number of burst releases		Low – All [p]'s, all [t]'s High – All [k]'s	BURST_NM
Voice onset time		Low – All [p]'s Med – All [t]'s, [ka], [ku] High – [ki]	VOT
F2 at voicing onset		Low – [pa], [pu], [ku] High – [pi], [ki], [ka], all [t]'s	F2_00MS
F2 10 msec after voicing onset		Low – [pa], [pu], [ku] High – [pi], [ki], [ka], all [t]'s	F2_10MS

Table 4. Summary of acoustic features predicted to cue listeners to stop place.

The feature variations by stop place suggest different roles in stop place identification for both humans and machines. The relative ranking and usefulness of each acoustic cue will be tested in further chapters.

- Spectral features were automatically extracted by linear and triangular fits to the spectrum at the burst. No single spectral shape feature is sufficient for correctly classifying the three stop places, but in combination these derived features capture the *diffuse-rising*, *diffuse-falling*, and *compact* classification that is characteristic of bilabial, alveolar, and velar stops, respectively.

- Relative amplitude is predicted to be most useful for detecting bilabial place, due to the similarity of velars and alveolars along this feature. Relative power, however, shows more variation by CV context.

- The number of bursts in the production of a stop is predicted to be most useful for distinguishing velar stops from other places of articulation.

- The semi-automatic extraction of VOT caused minimal measurement errors. VOT is expected to be a primary feature for stop place identification by both humans and machines.

- F2 onset features are predicted to be useful in detecting bilabials from other stop places, despite poor formant tracker estimates.

Now that the relevant features for stop place have been described, Chapter 3 will present a machine learning technique for estimating their usefulness to both humans and machines in stop place identification.

Chapter 3

Quantifying featural and contextual salience

1 Introduction

The previous chapter presented a set of acoustic features that were automatically extracted from the acoustic signal of 913 CV tokens. These features are all potential stop place cues for listeners. In the current chapter, *decision trees* (DTs), a machine learning approach to classification, are used to estimate the relative ranking of the inherent discriminatory power of these cues for each CV context and to show that this ranking varies by CV context.

Listener perception of stop place and the acoustic features that cue stop place are traditionally investigated by the systematic manipulation of one or two cues, often facilitated by the use of synthetic speech. The manipulated signal is presented to listeners, whose perception of stop place reveals the importance of the acoustic features involved. The current study attempts to test the relative ranking of the large set of cues suggested in previous phonetic studies and to show that this ranking varies by context. The multiple candidate cues to stop place discussed in the previous chapter are used in combination with DTs to establish a metric for quantifying each feature's potential contribution to stop place classification. Though stop place classification by automatic classifiers, such as DTs, are not models for human perception, they can nevertheless map out the relevant acoustic properties of a given segment with respect to others.

Section 2 briefly introduces DTs. The remainder of the chapter consists of DT classifications of stop place based on varying sets of acoustic features. In Section 3, DT classifications between a given consonant pair (for example, [pi] and [ti]) by a *single* acoustic feature yield an estimate of the discriminatory power of each acoustic feature. The same consonant pairs can also be classified by *multiple* acoustic features, as shown in Section 4. These results reveal feature interactions and can serve as a metric to the contextual salience of the acoustic signal by CV pair. In Section 5, DT classifications of *all three stops* using multiple acoustic features in a given vocalic context provide an overall estimate of the inherent informativeness of each vowel context. This last type of DT classification closely parallels the listeners' task in a classic perception experiment.

DT classifications of stop place reveal variations in the inherent informativeness of particular aspects of the acoustic signal by CV context, suggesting that differences in human perception errors by context may be due at least in part to differences in the relative informativeness of the signal in a given context. This hypothesis will be tested in Chapter 4 by a perceptual experiment.

2 Decision trees (DTs)

Decision trees are machine classifiers that represent their classifications in a readable *tree* form. The DTs used in the current study are CART-style (Breiman et al. 1984), as implemented by the IND package (Buntine & Caruana 1992). Much of the following summary is extracted from the IND package manual pages as well as from Russel & Norvig's (1995) textbook. Although any of several probabilistic classifiers could be used to classify stop place by acoustic features, DTs are particularly appropriate since they can

employ both discrete and continuous features. Additionally, they make no assumptions about the shape or scale of feature distributions, and the resulting tree form is highly interpretable (Figure 1).

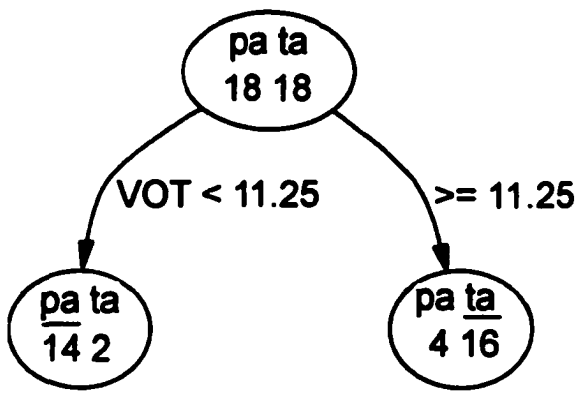


Figure 1. DT classification of [pa] and [ta] tokens by their VOT value. Nodes are represented by circles that contain classes and their associated number of tokens. The branches are indicated by the feature and associated threshold.

The tree structure of a DT represents a series of decisions that determine how a particular token or instance should be classified; these decisions correspond to a path through the tree from the *root node* (a node without parents) to a *leaf node* (i.e., one without any children). Each (non-leaf) node of a DT is associated with a decision or test, typically based on the value of a particular feature or set of features. The path taken in classifying a specific token depends on whether the test is satisfied, where the different child nodes correspond to the possible test conditions. In Figure 1, [pa] and [ta] tokens are tested for their VOT values. If a token has a VOT shorter than 11.25 msec, the token takes the left branch, if not, the token takes the right branch.

A DT is used to classify an instance by starting at the root node and following the path dictated by the feature tests until a leaf node is encountered. Each leaf node

corresponds to a classification(as indicated by an underline in Figure 1), and the instance is classified with the class assigned to that leaf node. In the current study, DTs are used to classify the place of articulation of CV stops by testing each CV token's value along a set of the semi-automatically extracted acoustic features described in Chapter 2.

When a DT is used as a learning classifier, the algorithm is said to perform a *supervised learning task*. The algorithm learns associations between classes and features in a *training set*, a portion of the data that is usually randomly selected. From these associations, the algorithm develops a classification rule to predict the class of further, unclassified examples (*test set*) based on their values along the features.

One standard technique for building classification rules from data is the *recursive partitioning algorithm* that forms the basis of the CART system (Breiman et al. 1984), used here. Recursive partitioning builds a tree from the root using the standard *greedy search* principle: At each node, always choose the feature test that results in the current best split and do not bother searching for a better one. The best node split is the one that most reduces the entropy of the data at that node (or provides the clearest class split). The tree may reuse a feature in more than one node test, as long as its reuse reduces entropy for the most data at that node. Once the training tree is built, the leaves of the tree store probabilities about the class distribution of all samples falling into the corresponding region of the feature space, which then serve as predictors for unseen test samples.

If the tree matches the training data too closely or if training data is not representative of test data, the DT may be unable to generalize to the unseen data. This is a particular risk when the training set is small, as it is in the current study where each CV context is trained on fewer than 60 individual tokens. The best DT is one that captures the

general trends of the data without *overfitting*, or matching the given data too closely, such that it will perform poorly when classifying test data. If a DT is thought to be overfitting the data, the tree may be modified after building by pruning back branches. *Cross-validation* is one technique for dealing with the problem of overfitting. During cross-validation, a fraction of the training set is set aside and used to test the prediction performance of a hypothesis induced from the remaining training data. This is repeated with different subsets of the training data, and results are averaged to form the training tree that will be used to classify the unseen data.

In the DT classifications of this thesis, the classes correspond to the three stops [p], [t], and [k] in a given context and data consists of the CV tokens and their associated values along the acoustic features discussed in Chapter 2. For 75% of the CV tokens (the training set), the DT induction uses the greedy search principle to learn DTs that capture associations between feature values and place of articulation. The DTs are then used to predict the place of articulation of the remaining CV tokens (the test set) based on their values along the same one or more acoustic features. The DT is pruned using a specific case of cross-validation, useful for small amounts of data, where the training set is randomly divided into two subsets of roughly equal size. The results from training on the two trees independently are averaged to form the final training tree. Specific examples are provided in each of the following sections.

3 Single-feature DT classification by consonant pair

In this section, I present DTs that distinguish between pairs of stops in a given vocalic context based on a single acoustic feature from the set of semi-automatically extracted features discussed in Chapter 2. The accuracy with which these single-feature DTs classify the relevant CV tokens in the test set provides a relative metric for the inherent discriminability of the given feature for the given CV pair.

In single-feature DT classification, the training set is used to establish one or more thresholds that split the tokens into subsets that maximize the number of training tokens correctly classified. The resulting DT is then used to classify the remainder of the test data.

Figure 2 shows the classification rule, or training tree, for distinguishing [p] and [t] in the context of [a] by the acoustic feature VOT. The DT was fed the set of 76 [pa] and 73 [ta] tokens, along with their values for VOT. In order to equalize the set sizes, the set of [pa] tokens was first *downsampled* to 73. The training set was then formed by randomly selecting 57 [pa] tokens and 57 [ta] tokens. The top node of the tree shows the number of [pa] and [ta] tokens included in the training set. For the purposes of this study, the set sizes are equalized, though the same DT could be run with set sizes that reflect the frequency of each CV token in the lexicon.

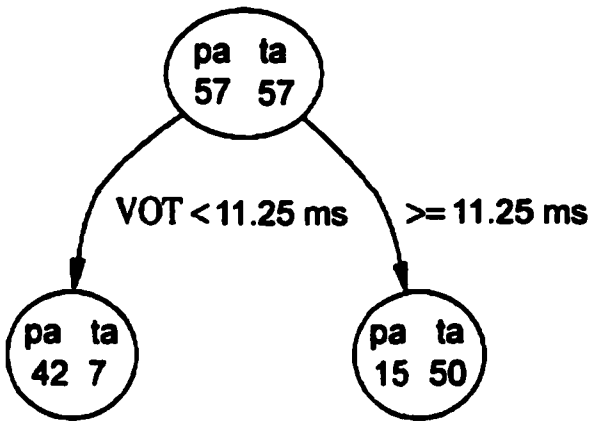
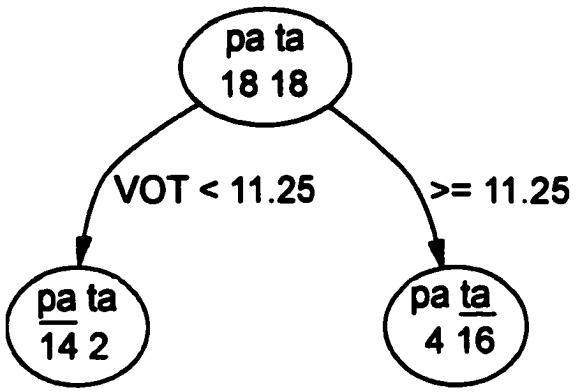


Figure 2. Training tree of [p] and [t] tokens in the context of [a] by the acoustic feature VOT. Nodes are represented by circles which contain classes and their associated number of tokens. The branches are indicated by the feature and associated threshold.

The tree develops the classification rule seen in Figure 2 by selecting the VOT threshold that maximally reduces the entropy of [pa] and [ta] tokens. The majority of [pa] tokens (42 out of 57) in the training set have VOT values shorter than 11.25 msec. Test CV tokens with VOTs shorter than this threshold will be classified as *pa* tokens. The majority of [ta] tokens (50 out of 57) in the training set have VOT values longer or equal to 11.25 msec. Test CV tokens with VOT values greater than 11.25 msec will be classified by the tree as *ta* tokens. The classification of the training data need not be perfect; in this case, 7 [ta] training tokens have VOT values shorter than 11.25 msec and 15 [pa] training tokens have VOT values longer than 11.25 msec, yielding an overall accuracy of 80.7%.

Finally, the test 18 [pa] and 18 [ta] tokens in the test set are classified along the branches formed by the training tree (Figure 3). Based on the classification determined by the training tree, 14 of the 18 test [pa] tokens have a VOT less than 11.25 msec and are

correctly classified. However, 4 of the 18 test [pa] tokens have VOT durations longer than 11.25 msec, resulting in their misclassification by the decision tree. Similarly, 16 of the 18 test [ta] tokens are correctly classified, and 2 are incorrectly classified. The overall accuracy of the classification is 83.33%, derived from the number of correctly classified [pa] and [ta] tokens (30) divided by the total number of test tokens (36).



Accuracy: 83.33%

Figure 3. DT classification of the test [pa] and [ta] tokens by VOT, using the training tree in Figure 2. Tokens in leaf nodes are assigned to the class of the underlined CV of that node.

The DT classifies [pa] and [ta] tokens with an accuracy of 83.33%, indicating that [pa] and [ta] tokens differ consistently by their VOT. Now let us turn to the DT classification of the same CV pair by the feature BURST_NM (number of bursts), which is predicted to be a useful feature for the classification of velars, but not for [pa] and [ta] tokens, since both bilabial and alveolar stops tend to have single stop bursts.

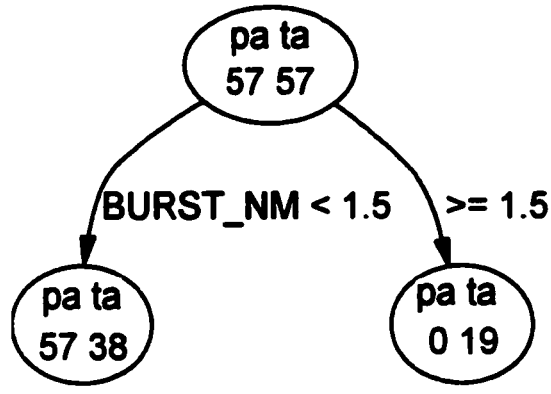
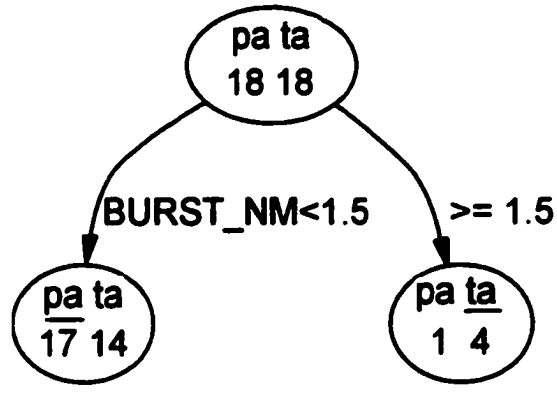


Figure 4. Training tree of [pa] and [ta] tokens by BURST_NM. Nodes are represented by circles which contain classes and their associated number of tokens. The branches are indicated by the feature and associated threshold.

The best classification for the [pa] and [ta] tokens in the training set using BURST_NM separates those tokens with a single burst from those with multiple bursts (Figure 4). Recall that the feature BURST_NM takes only whole integer values from 1 to 4. 19 of the 57 training [ta] tokens have multiple bursts. The remainder of the [ta] tokens and all [pa] tokens in the training set have a single burst only. Classification of the test data is predicted to be worse for BURST_NM than for VOT, based on the number of tokens misclassified by the training tree.

As predicted, the classification of the test tokens using the feature BURST_NM performs only slightly above chance (58.33%). Although the majority of the [pa] tokens are correctly classified, only 4 [ta] tokens are. The low accuracy rating for this feature indicates that the number of bursts in a stop release is not inherently informative for the discrimination of [pa] and [ta], though it might be informative for other pairs of stop consonants, such as pairs that include velar stops. The remainder of this section reports

on the performance of single-feature DTs based on each of the acoustic features available in the dataset, grouped by vocalic context.



Accuracy: 58.33%

Figure 5. DT classification of test [pa] and [ta] tokens by BURST_NM, using the training tree in Figure 4.

Each DT is pruned using cross-validation between two subsets of the training data. The thresholds from training on the two subsets independently are averaged to form the final training tree. Specifics on the test and training size will be given for each classification. Downsampling was used in each case to equalize the number of tokens across stop place categories. The accuracy rate of single-feature DT classifications are shown by stop pair in each of the three vowel contexts. A high percentage of correctly classified CV tokens by a given feature indicates a high level of inherent informativeness of that feature between a given pair of CVs.

3.1 DT classification of [p] and [t]

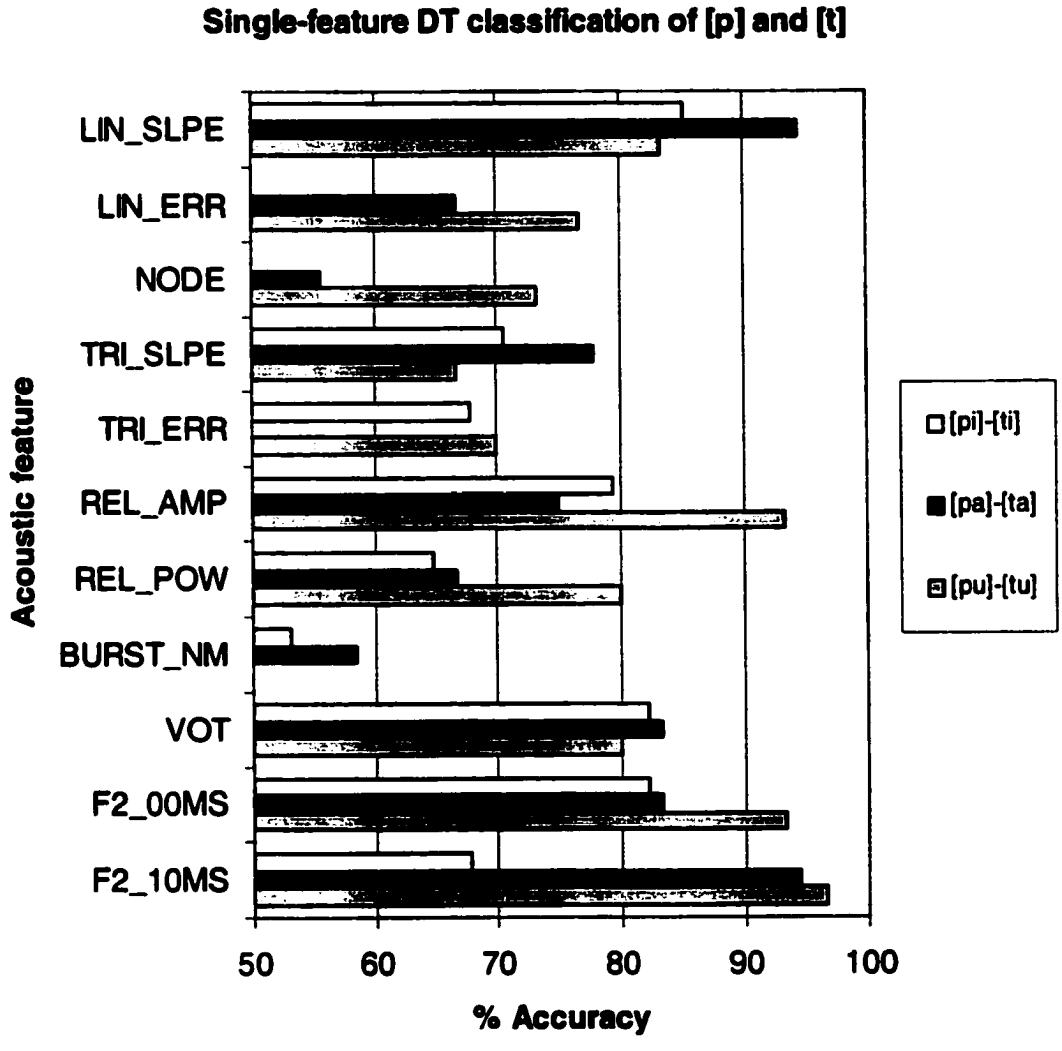


Figure 6. Single-feature DT classifications for [p] and [t] tokens by vocalic context. The percentage of correctly classified tokens is reported above chance only.

DTs for [pi] / [ti] classification were trained on 110 CV tokens (55 [pi] and 55 [ti]) and tested on 34 CV tokens (17 [pi] and 17 [ti]). DTs for [pa] / [ta] classifications were trained on 114 CV tokens (57 [pa] and 57 [ta]) and tested on 36 CV tokens (18 [pa] and

18 [ta]). DTs for [pu] / [tu] classification were trained on 120 CV tokens (60 [pu] and 60 [tu]) and tested on 30 CV tokens (15 [pu] and 15 [tu]).

In general, bilabials and alveolars are best differentiated by LIN_SLPE and F2 (F2_00MS, F2_10MS) (Figure 6). The remaining spectral features, REL_POW, and BURST_NM are inherently uninformative for bilabial/alveolar distinctions across all vowel contexts.

The inherent discriminibility of acoustic features between bilabial and alveolar stop place varies greatly by vocalic context. For example, alveolar and bilabial stops in the contexts of [a] and [u] can be correctly classified at 85% accuracy or more by more than one feature. No acoustic feature performs at this level when the vocalic context is [i]. For example, DT classifications of stop place using only F2 at 10 msec after voicing onset (F2_10MS) perform with near perfect accuracy for contexts [a] and [u], but only at an accuracy of only 68% when the context is [i]. This supports the prediction that for bilabials and alveolars in the context of high front vowels, there is less information about the place of articulation available in the signal for listeners than in other vocalic environments.

3.2 DT classification of [t] and [k]

DTs for the [ti] / [ki] classification were trained on 110 CV tokens (55 [ti] and 55 [ki]) and tested on 34 CV tokens (17 [ti] and 17 [ki]). DTs for [ta] / [ka] classification were trained on 114 CV tokens (57 [ta] and 57 [ka]) and tested on 32 CV tokens (16 [ta] and 16 [ka]). DTs for [tu] / [ku] classification were trained on 114 CV tokens (57 [tu] and 57 [ku]) and tested on 30 CV tokens (15 [tu] and 15 [ku]).

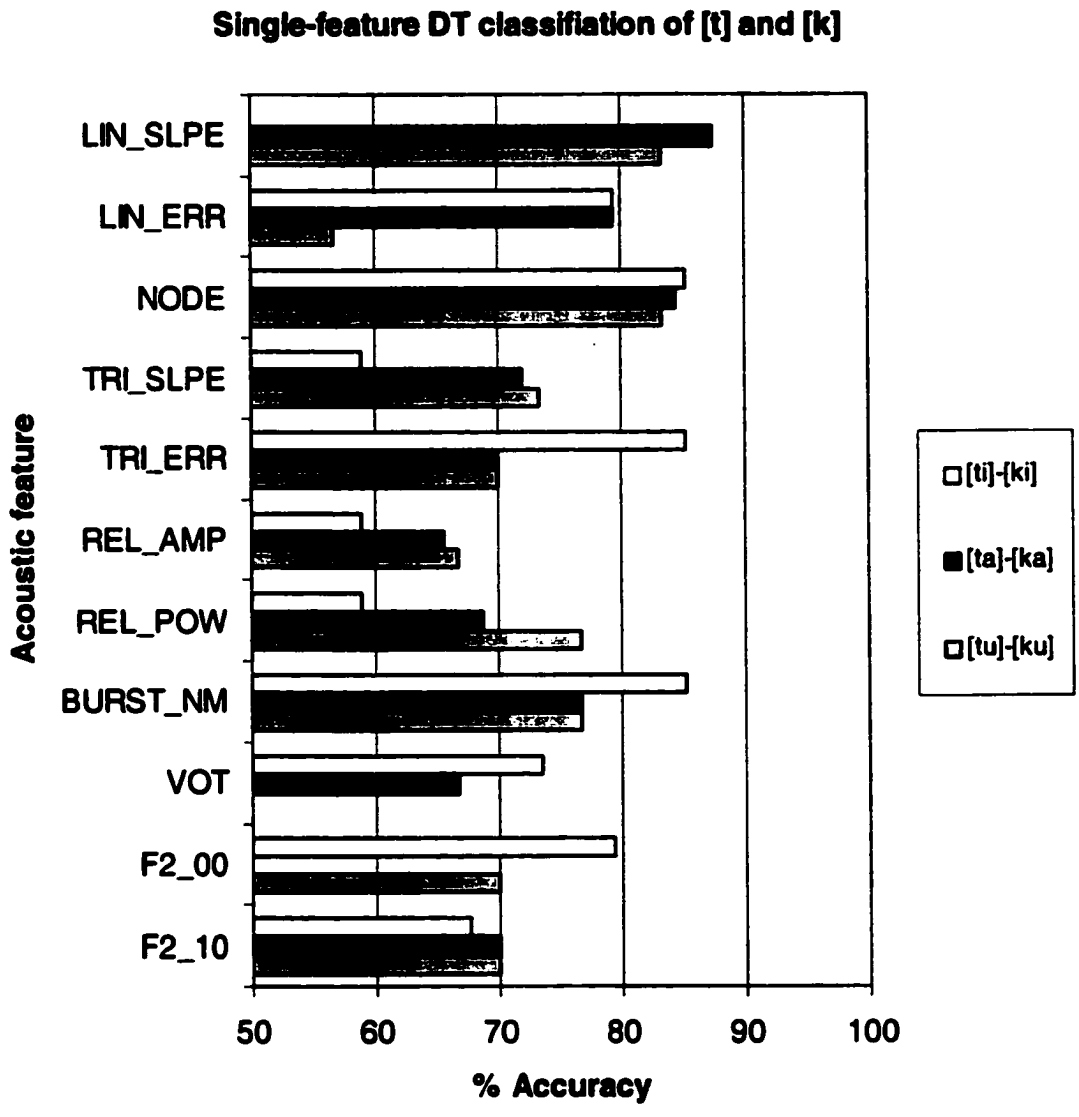


Figure 7. Single-feature DT classifications for [t] and [k] tokens by vocalic context. The percentage of correctly classified tokens is reported above chance only. See Chapter 2, Table 4 for a key to the feature abbreviations.

For single-feature DT classifications of stops [t] and [k] (Figure 7), most accuracy rates are lower than in the classification of [p] and [t] (Figure 6). Alveolar and velar stops are therefore more similar along the set of acoustic features studied here than bilabial and alveolar stops. Assuming that the set of acoustic features is comprehensive, there is less information available in the signal for the discrimination of [t] and [k] than

for [p] and [t]. The node of the triangular fit to the burst spectrum (NODE) distinguishes alveolars from velars with a relatively high accuracy (~84%) for all three contexts, supporting the prediction that the spectral peak at the burst may contain sufficient information to classify alveolars from velars regardless of vocalic context.

In alveolar and velar stop place discrimination, the accuracies of DT classifications by each feature vary greatly by the following vowel. In particular, in the [i] context, the slope of the linear fit to the burst spectrum (LIN_SLPE) is too similar for [t] and [k] tokens to classify stop place at better than chance, whereas in the [a] and [u] contexts, classification along this feature yields the highest accuracy for all acoustic features shown. Features TRI_ERR and BURST_NM correctly classified alveolar and velar stops in the [i] context at 85% accuracy, though in the [a] context they made many more errors in stop place classification.

3.3 DT classification of [k] and [p]

DTs for the [ki] / [pi] classification were trained on 120 CV tokens (60 [ki] and 60 [pi]) and tested on 36 CV tokens (18 [ki] and 18 [pi]). DTs for [ka] / [pa] classifications were trained on 120 CV tokens (60 [ka] and 60 [pa]) and tested on 36 CV tokens (18 [ka] and 18 [pa]). DTs for [ku] / [pu] classification were trained on 114 CV tokens (57 [ku] and 57 [pu]) and tested on 32 CV tokens (16 [ku] and 16 [pu]).

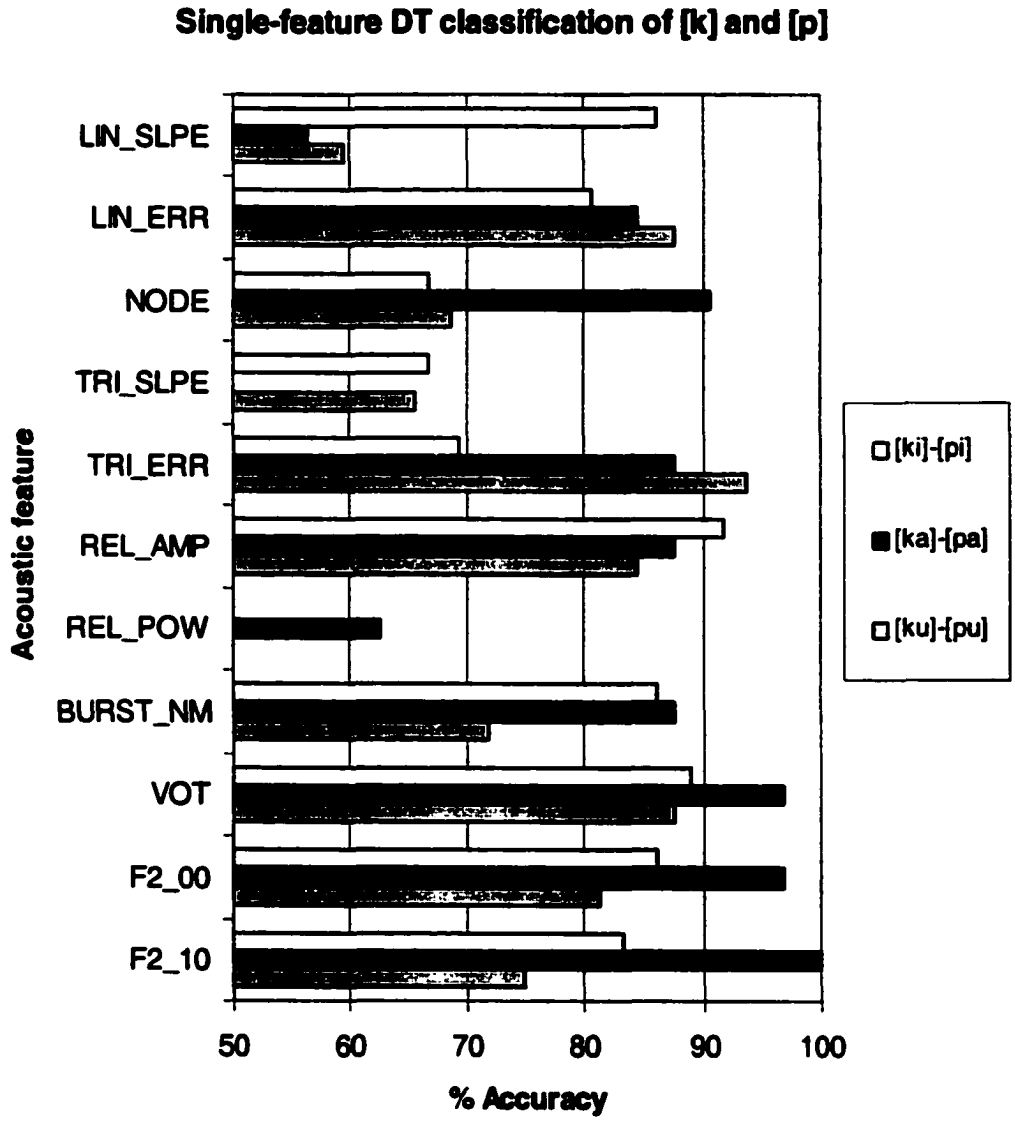


Figure 8. Single-feature DT classifications for [k] and [p] tokens by vocalic context. The percentage of correctly classified tokens is reported above chance only. See Chapter 2, Table 4 for a key to the feature abbreviations.

Single-feature DT classifications of stops [p] and [k] yield high accuracy rates for the numerous acoustic features across all vowel contexts (Figure 8). Stops [p] and [k] differ systematically along the majority of features, making them easier to classify by

DTs than both [p] / [t] and [t] / [k] pairs. [p] and [k] in the [u] context are classified with high accuracy rates for LIN_ERR, TRI_ERR, REL_AMP, and VOT. Although the [i] context produces the lowest accuracy ratings of the three vowel contexts, four acoustic features perform at 85% or higher when classifying [pi] and [ki]. The single-feature DT classifications of [k] and [p] indicate an acoustic signal rich in discriminatory place information, especially in the [a] context. For all three vowel contexts, the acoustic features that provide discriminations higher than 85%, for example, are from different portions of the acoustic signal: properties of the burst spectrum, formant transitions, relative amplitude, and VOT. Indeed, across all vocalic contexts, only two acoustic features, REL_POW and TRI_ERR, perform at low accuracy rates for [p] / [k] discrimination.

3.4 Summary and discussion

The performance of acoustic features in DT classification by CV pair varies by vowel context and by stop pair, suggesting that the portion of the acoustic signal most relevant for listener perception of stop place varies depending on the CV context. In general, the accuracies of DT stop place classifications are lower in the [i] context than in other vocalic contexts. This supports the prediction that for stops in the context of high front vowels, there is less information about the place of articulation available in the signal for listeners to use than in other vocalic environments.

Specifically, bilabial and alveolar stops are classified with the highest accuracies by the LIN_SLPE and F2 features when the following vowel is an [a] or [u]. The stops [t] and [k] have the least amount of differentiating information available in the acoustic

signal for listener perception of stop place of all CV pairs. The feature NODE (the node of the triangular fit to the burst spectrum), however, distinguishes alveolars from velars with a relatively high accuracy (~84%) for all three vocalic contexts, supporting the prediction that the spectral peak at the burst may contain sufficient information to classify alveolars from velars regardless of the following vowel. Stops [p] and [k] differ systematically along a majority of features from different portions of the acoustic signal, especially LIN_ERR, TRI_ERR, REL_AMP, and VOT. The single-feature DT classifications of [k] and [p] indicate an acoustic signal rich in discriminatory place information, especially in the [a] context.

4 Multiple-feature DT classification by consonant pair

Human identification of stop consonant place is clearly more complex than single-feature DT classifications between CV pairs. Listeners are known to rely on many acoustic cues at once in phoneme identification tasks. Many perception studies have found interacting effects between acoustic features, such as spectral and formant information at the onset of the following vowel (Cooper et al. 1955; Dorman & Loizou 1996). Interaction effects can be studied only when multiple features are considered in a given discrimination task. In the following section, I present the results from DT classifications of CV pairs that use *all* available acoustic features. Multiple-feature DT classification permits the investigation of complex interactions among acoustic properties of the signal.

Additionally, many of the acoustic features studied here represent properties of the signal that, when considered individually, do not correspond to cues used by listeners. For example, previous perceptual studies (Blumstein & Stevens 1978; Stevens &

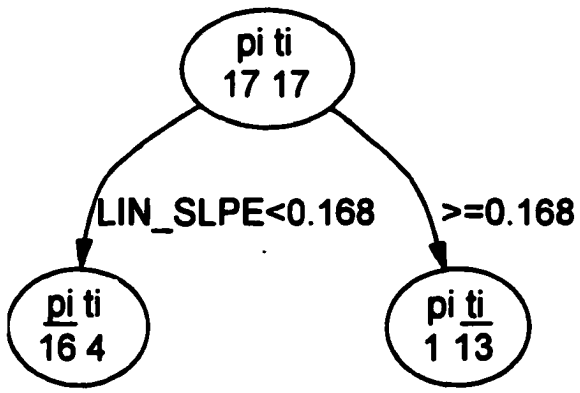
Blumstein 1979) indicated that listeners are sensitive to the gross shape of the spectrum at the burst, associating a diffuse-falling shape with bilabials, a diffuse-rising shape with alveolars, and a mid-compact shape with velars. In the current study, the set of spectral features derived from a linear and triangular fit to the spectrum (LIN_SLPE, LIN_ERR, TRI_SLPE, TRI_ERR, NODE) must be used in combination to capture this three-way classification.

The DT learning algorithm for multiple features is the same as in the case of single-feature DT classification, except that the training tree may continue to split at each subnode if such a split reduces the overall entropy. For most cases, the training tree will only use one or two acoustic features. The algorithm stops splitting when no possible feature test reduces entropy.

As mentioned in Section 2, one drawback to the greedy search principle employed by the DT algorithm is that a root node feature test may split the data in a way that causes suboptimal classification overall, although it is the best split at that point (in terms of entropy reduction). To avoid this problem, DTs with multiple branches were run several times, first with the top feature omitted from the available acoustic features, then with the second feature omitted, and so on. In each case, the tree with the highest overall accuracy is considered the best fit to the data and is presented in this section.

In the following trees, the DT is pruned using cross-validation between two subsets of the training data. The results from training on the two trees independently are averaged to form the final model for the training tree. Specifics on the test and training size will be given for each classification. Downsampling was used in each case to equalize the number of tokens across stop place categories.

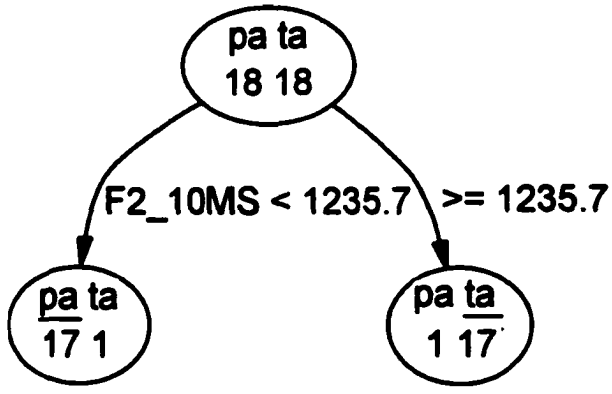
4.1 DT classification of [p] and [t]



Accuracy: 85.29 %

Figure 9. Multiple-feature DT classification of test [pi] and [ti] tokens. Only LIN_SLPE is used by the DT for classification. The DT is trained on 110 tokens (55 [pi] and 55 [ti]).

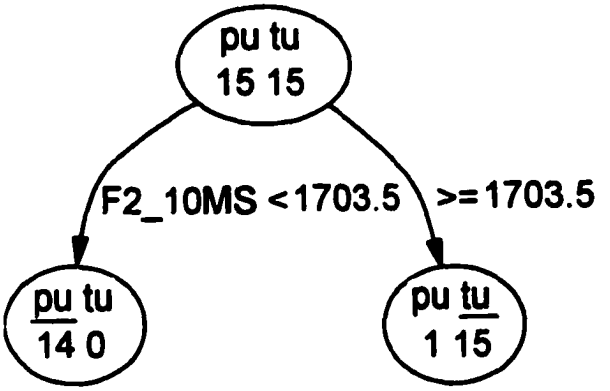
The DT training algorithm for classifying [pi] and [ti] tokens was performed with all acoustic features available, but only the slope of the linear fit to the burst spectrum (LIN_SLPE) was used by the DT for classification (Figure 9). Recall that LIN_SLPE classified [pi] and [ti] with the highest accuracy in the single-feature DT classifications, with VOT following (Figure 6). Therefore, no combination of features classifies [pi] and [ti] better than the best single-feature classification. The overall accuracy of the [pi] / [ti] classification by LIN_SLPE is only 85.29%, which is much lower than other CV pairs presented in this section. The relatively low accuracy indicates that the acoustic signal has less discriminatory information available for distinguishing between [p] and [t] in the [i] context than in other contexts.



Accuracy: 94.44%

Figure 10. Multiple-feature DT classification of test [pa] and [ta] tokens. Only the feature F2_10MS is used by the DT for classification. The DT is trained on 114 CV tokens (57 [pa] and 57 [ta] tokens).

The stop consonants [p] and [t] in the context of [a] are classified with an accuracy rating of 94.44%, using only their values of F2 extracted at 10 msec after voicing onset (F2_10MS) (Figure 10). All other features were available for this classification, but no combination of features added extra power in entropy reduction. When the feature F2_10MS is removed and the DT rerun, the feature LIN_SLPE fares equally well at classifying [p] and [t] (94.44%) by splitting the data at the threshold 0.117. (All but one [p] and one [t] token are correctly classified by LIN_SLPE.) Recall that in single-feature DT classifications, LIN_SLPE and F2_10MS are the most discriminatory features in bilabial/alveolar discrimination. [p] / [t] discrimination in the [a] context is possible because of not one but two highly informative features.



Accuracy: 96.67%

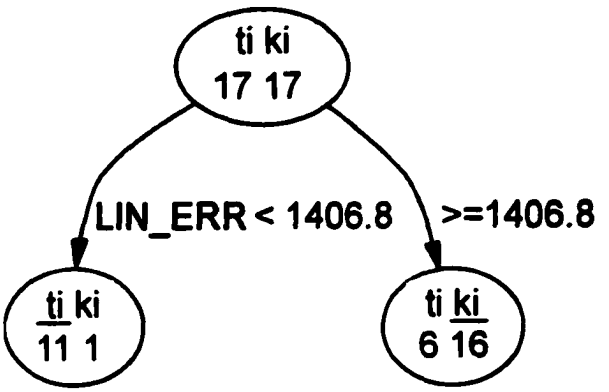
Figure 11. Multiple-feature DT classification of test [pu] and [tu] tokens. Only the feature F2_10MS is used by the DT for classification. The DT is trained on 120 CV tokens (60 [pu] and 60 [tu]).

For the [u], as well as the [a] context, the value of F2 at 10 msec after voicing onset (F2_10MS) is used exclusively by the DT to classify [pu] and [tu] with 96.67% accuracy (Figure 11). Only one bilabial with an uncharacteristically high F2 onset was misclassified as an alveolar by the DT. The threshold of the DT classification (1703.5 Hz), as compared to that of the [pa] / [ta] classification (1235 Hz), reflects the higher F2 values (~1800 Hz) of alveolar stops in the [u] environment than in the [a] environment. Bilabials in both the [a] and [u] environment have F2_10MS values around 1100 Hz.

Once again, though all acoustic features were available, only one highly informative feature is used by the learned DT, indicating that no combination of features fares better in the classification of bilabials and alveolars. Recall, however, that REL_AMP and F2_00MS are also highly informative features for this CV pair; their single-feature DT classifications both perform at 94.44% accuracy.

4.2 DT classification of [t] and [k]

Unlike DT classifiers for [p] and [t], which use only a single highly informative feature, many of the DT classifiers for [t] and [k] involve multiple acoustic features. In general, the accuracy rates are lower for [t] / [k] classification than for [p] / [t] classification. Once again, stop place classification in the [i] context yields the lowest accuracy rate of all three vocalic contexts.

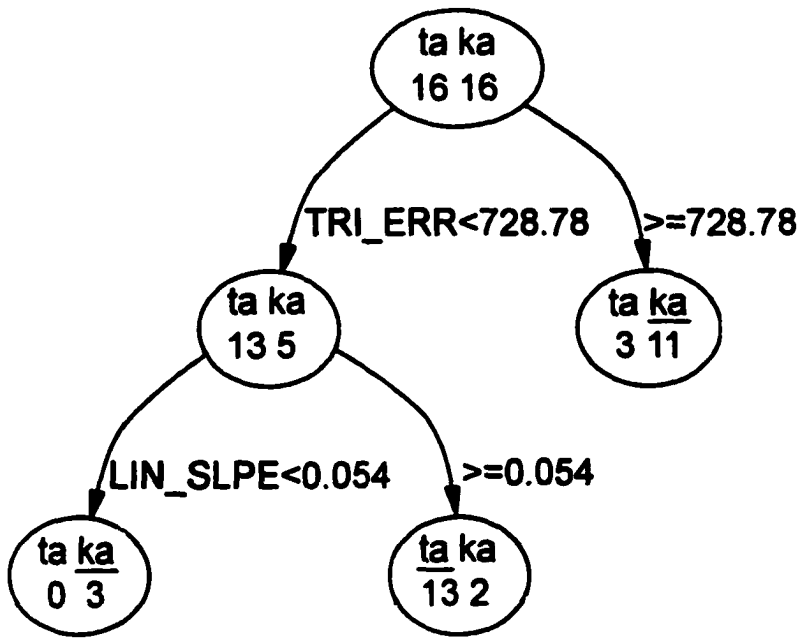


Accuracy: 79.41%

Figure 12. Multiple-feature DT classification of test [ti] and [ki] tokens. Only the feature LIN_ERR is used by the DT for classification. The DT is trained on 110 tokens (55 [ti] and 55 [ki] tokens).

All acoustic features were available for the classification of [ti] and [ki] tokens. Only the feature LIN_ERR is used in the resulting DT; lower errors of the linear fit to the burst spectrum tend to correspond to [ti] tokens, while higher errors capture the mid-frequency peak of the [ki] tokens (Figure 12). The overall accuracy of this tree is low as compared to other contexts, suggesting that in the [i] context, less discriminatory information about the stop place is available in the acoustic signal.

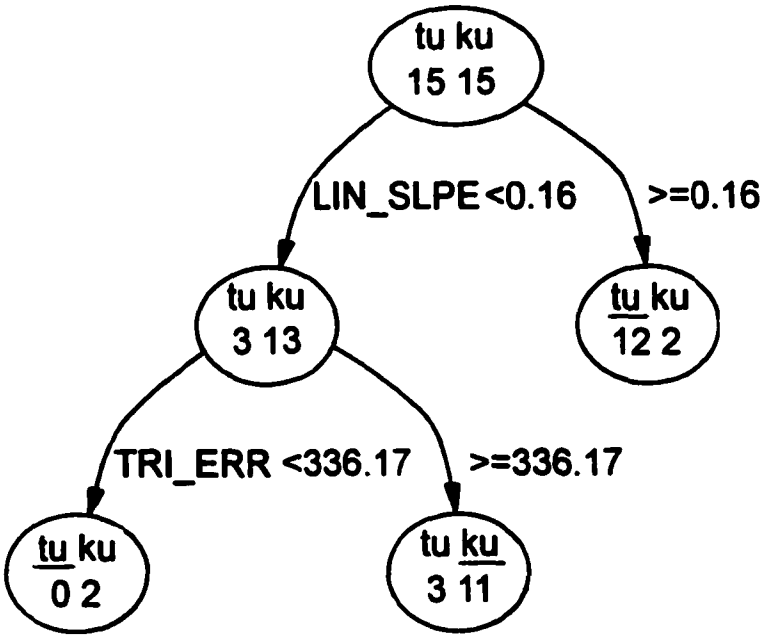
In the DT classification of [ta] and [ka] tokens, the acoustic features TRI_ERR and LIN_SLPE interact to classify place of articulation (Figure 13). Stops with high error of the triangular fit to the burst spectrum (TRI_ERR) are classified as velars, capturing their characteristic mid-frequency peak. Stops with low TRI_ERR values are further split by the slope of the linear fit to their burst spectrum (LIN_SLPE); alveolars have a steep slope and velars have a flatter slope (cf. Chapter 2). The DT classification of [ta] and [ka] tokens demonstrates how spectral fit features work together to group stops by the overall shape of their burst spectrum. The accuracy with which these features classify [ta] and [ka] is higher than in the corresponding [i] context, but lower than classification between [pa] and [ta], for example.



Accuracy: 84.38%

Figure 13. Multiple-feature DT classification of test [ta] and [ka] tokens. The DT was trained on 114 CV tokens (57 [ta] and 57 [ka] tokens).

DT classifiers for [t] and [k] in the context of [u] have the lowest accuracy rate of all pairwise classifications: 76.67% (Figure 14). The trees first distinguish [tu] from [ku] by the slope of the linear fit to the burst spectrum (LIN_SLPE); characterizing alveolars by their steep spectral tilt at the burst. Among the set of training CVs, the alveolars with uncharacteristically flat spectral tilts are distinguished from velars by their low errors to the triangular fit to the spectrum (TRI_ERR). In the particular set of test [tu] and [ku] tokens shown in Figure 14, however, all three of the [tu] tokens have high errors to the triangular fit and are thus misclassified as velars by the tree. Similarly, two [ku] tokens are misclassified for their low error to the triangular fit at the burst. These misclassifications contribute to the relatively low accuracy rating of the DT classification.



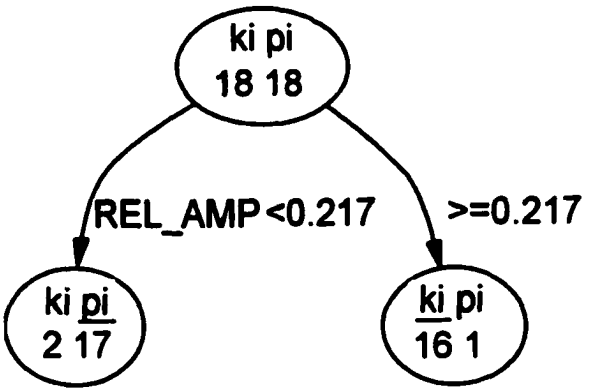
Accuracy: 76.67%

Figure 14. Multiple-feature DT classification of test [tu] and [ku] tokens. The DT is trained on 114 CV tokens (57 [tu] and 57 [ku]).

A second DT was run with a new randomly selected test and training set for [tu] and [ku] tokens to determine whether the misclassifications could be attributed to the particular division of data chosen, or whether [tu] and [ku] have inherently less discriminatory information available for place distinctions in the signal. The resulting DT (not shown) employs the same features with slightly different thresholds. In particular, TRI_ERR divides alveolars from velars at 608.76 (instead of 336.17), causing fewer misclassifications in the two resulting nodes. A greater number of tokens misclassified in the other leaf node, however, results in a comparable accuracy rate, indicating that the low accuracy is due to acoustic similarity between [tu] and [ku].

4.3 DT classification of [k] and [p]

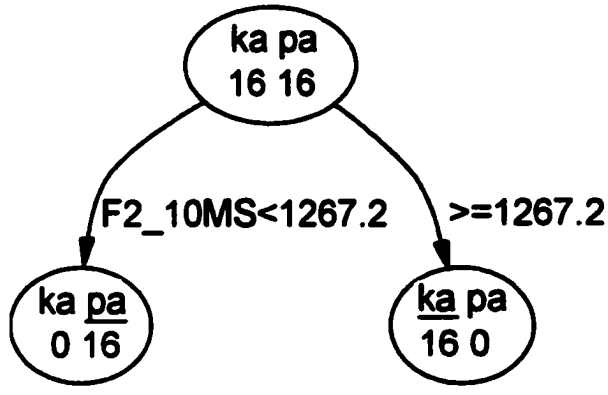
Compared with DT classifications of other stop pairs ([p] / [t] and [t] / [k]), DT classifiers of [k] and [p] stops in all environments have the highest accuracy rates. Though some errors in DT classification may occur, the acoustic features extracted from the signal in the current study provide ample discriminatory information for [k] and [p] classification. The primary acoustic feature or features employed by the DT for stop place classification varies by vocalic context.



Accuracy: 91.67%

Figure 15. Multiple-feature DT classification of test [ki] and [pi] tokens. The DT uses only REL_AMP for classification. The DT is trained on 120 tokens (60 [ki] and 60 [pi]).

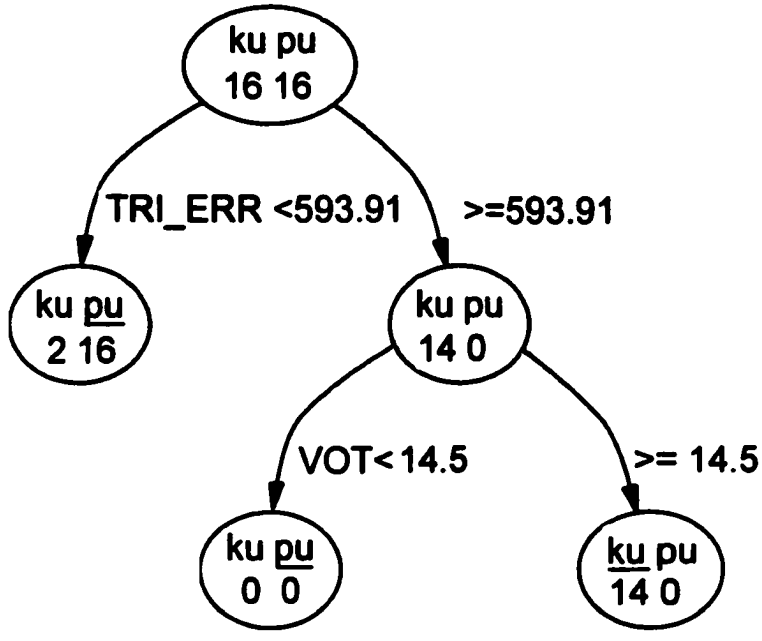
In the DTs seen so far, the [i] environment is associated with the lowest accuracy rates for DT classification of consonant pairs. DT classification of [ki] and [pi] performs at 91.67%, using only the relative amplitude of the burst with respect to the following [i] (REL_AMP) (Figure 15). Though 91.67% is a lower accuracy than in the other [p] / [k] DT classifications, it is still higher than for most CV pairs. Note that other features, such as VOT, also perform at high accuracy rates for this pair. The acoustic signal thus seems to be rich in information for the discrimination between stops [p] and [k] in the environment of [i].



Accuracy: 100%

Figure 16. Multiple-feature DT classification of test [ka] and [pa] tokens. The DT uses only F2_10MS for classification. The DT is trained on 120 CV tokens (60 [ka] and 60 [pa]).

A DT using the same feature as in the [pa] / [ta] DT classifier, F2 at 10 msec after voicing onset (F2_10MS), classifies [pa] tokens from [ka] tokens with 100% accuracy. Further investigation of the distribution of tokens along their F2_10MS values (Chapter 5) shows that the [pa] and [ka] tokens in this database do not overlap at all along this feature, resulting in perfect DT classification. Sufficient information for distinguishing between [pa] and [ka] is available in the second formant alone for both machines and listeners.



Accuracy: 93.75%

Figure 16. Multiple-feature DT classification of test [ku] and [pu] tokens. The DT is trained on 114 CV tokens (57 [ku] and 57 [pu]).

Labials and velars are accurately classified in the context of [u] by the error of the triangular fit to their burst spectrum (TRI_ERR) and by their VOT (Figure 16). Bilabials have characteristically flat burst spectra as compared to the prominent mid-frequency peak of velar stops and so are classified by their low values along TRI_ERR. Those labials with uncharacteristically high errors to the triangular fit are further classified by their relatively short VOT, though in this particular test set, no CV tokens satisfy this criteria.

4.4 Summary and discussion

Multiple-feature DT classifications of CV pairs reveal variations in the inherent informativeness of particular aspects of the signal by CV context, suggesting that

differences in human perception errors by context may at least in part be due to differences in the relative informativeness of the signal in a given context.

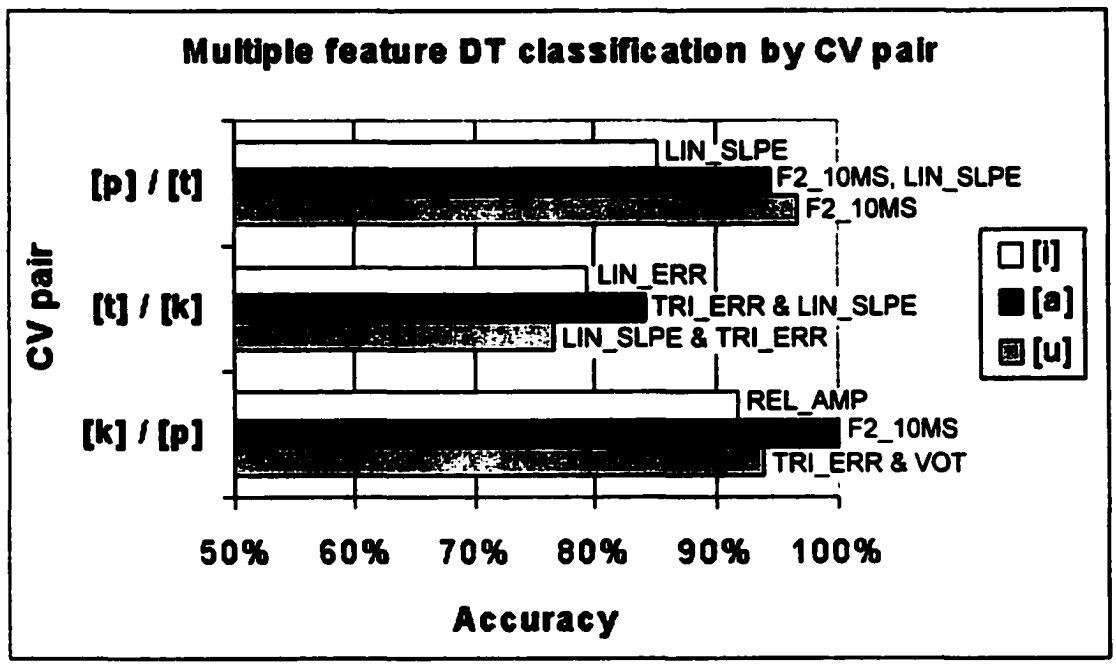


Figure 17. Summary of the accuracy rates of multiple-feature DT classification of stop pairs. The feature or features used by the DT are listed to the right of the bar.

The multiple-feature DT classifications of [t] and [k] had the lowest accuracy rates of all CV pairs. In all three vocalic contexts, spectral features yielded the best classification of the [t] and [k] tokens, suggesting that [t] and [k] are quite similar along the other acoustic features studied. The differences between [t] and [k] in the spectrum at the burst appear to be the most consistent difference in the acoustic signal relevant to stop place, but they are classify only 80% of test tokens accurately.

The [i] context produces DTs with lower accuracy than the [a] context for all three pairs. The accuracy for the [i] context is also lower than the [u] context for both [p]

/ [t] and [k] / [p] classifications. In addition, DT stop classification in the [i] context never relies on formant features such as F2_00MS or F2_10MS. The high degree of error for machine classification of stops in the [i] context suggests that problems in listener identification of stop place are due to an impoverished signal in the context of [i], in particular, the neutralization of F2 onset information by the high F2 of [i]. Stop classification of bilabial, alveolar, and velar stops in the [i] context must rely on spectral properties, temporal properties, or relative amplitude for discrimination along stop place.

With the exception of [pa] / [ta] classification, the [a] context usually yields classification accuracies that are slightly higher than the corresponding [u] context. [pa] tokens, for example, are classified with little error by their low F2 onset in DT classifications with both [ta] and [ka].

[ku] is distinguished from [pu] and [tu] by the feature TRI_ERR in combination with another feature in both cases. TRI_ERR is high in the case of [ku], capturing not only the characteristic mid-frequency peak of the velar burst, but also the secondary peak in the lower frequencies created by the following rounded vowel, [u].

[pu], like [pa], is characterized by a low F2 onset, but this feature is not the primary feature for DT classification of [pu] and [ku]. In general, velars share the property of low F2 onsets with bilabials, motivating the Jakobsonian feature *[+grave]*, defined as segments having predominately low-frequency energy (Jakobson et al. 1952), which includes labials, velars, labiovelars, and rounded vowels.

The variations in feature usage and corresponding accuracy rates by vocalic contexts provide an insight into variations in the cues available to listeners for identifying particular stops. If listeners identify stop place by using a combination of the most

informative acoustic features for a given CV context, they are expected to perform well in contexts in which there are many highly informative cues and to perform poorly in contexts in which the cues are inherently less informative. Although the exact mechanisms of human stop perception are much more complex than the DT classifiers shown here and the extraction method adopted in this thesis may not have captured all acoustic cues, listeners can only do as well as the signal allows. A less informative signal is predicted to cause more listener error than a signal rich with discriminatory information. This hypothesis will be tested in Chapter 4, in which listener confusions in an experimental perception study are compared to the relative trends of the multiple-feature DT classifiers described here.

5 Multiple-feature DT classification by vocalic context

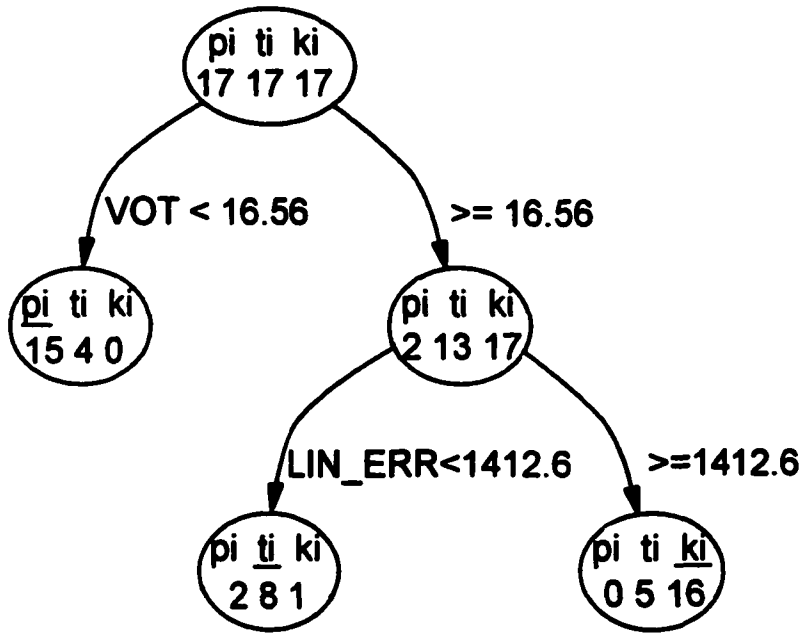
The performance of pairwise DT classifiers provides an estimate of the inherent discriminatory power of the acoustic signal for a given context, but DTs do not mimic the identification task that human listeners perform in the majority of phoneme identification studies. In the majority of listener tasks involving CV stop place identification, the listener identifies the place of an initial stop consonant from a set of more than two options; usually the choices are bilabial, alveolar, and velar stop place. In some cases the following vowel of the CV stimuli is fixed. In others, CV stimuli are presented such that the following vowel changes randomly. In this section, multiple-feature DT classifiers are induced for *all three stops* in each of the vowel contexts. The performance of these DT classifiers reveals variations by vocalic context that mirror classic perceptual studies in which the identity of the following vowel is known by the listener. Finally, the results

from a multiple-feature DT classification of all three stop places is presented across all CV tokens regardless of the identity of the following vowel. This final classification mimics human perceptual studies in which CV tokens for all three vowels are presented randomly to listeners for stop place identification.

As in the previous cases, the trees described below are pruned using cross-validation between two subsets of the training data. The results from training on the two trees independently are averaged to form the final model for the training tree. Specifics on the test and training size will be given for each classification. Downsampling was used in each case to equalize the number of tokens across stop place categories.

5.1 DT classification of stop place in the [i] environment

In the DT classification of [pi], [ti], and [ki] tokens with all acoustic features available, the [pi] tokens are best distinguished from other stop places by their low VOT values (Figure 18). The majority of both [ti] and [ki] tokens have longer VOTs than the bilabial stops in this context. The DT further splits [ti] and [ki] tokens by their LIN_ERR values; C[i] tokens with large LIN_ERR are classified as *ki*, which accurately captures the majority of [ki] tokens. Alveolar tokens are defined as those tokens with a long VOT and a small LIN_ERR.



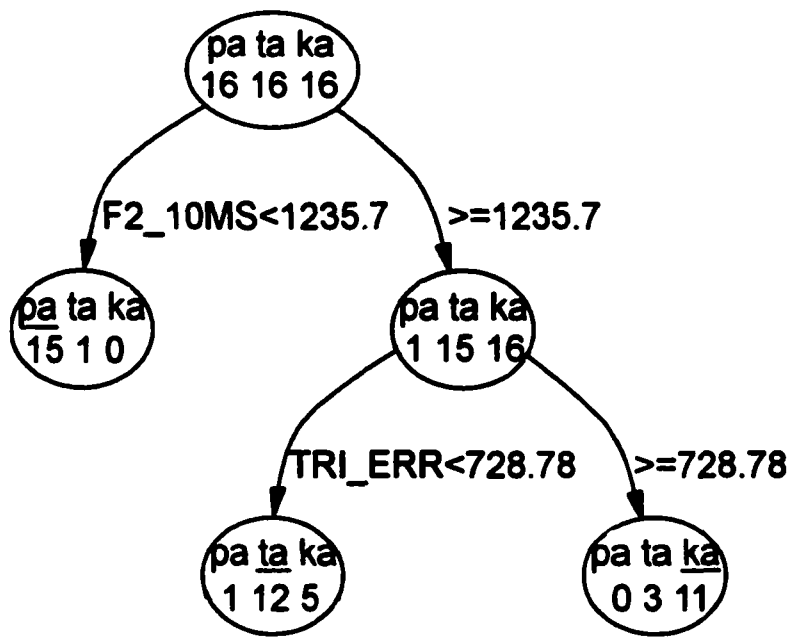
Accuracy: 76.47%

Figure 18. DT classification of test [pi], [ti], and [ki] tokens. The DT is trained on 165 CV tokens (55 [pi], 55 [ti], and 55 [ki]).

In the test set shown here, [ti] tokens are misclassified by the tree structure far more than [pi] or [ki] tokens. This is opposite of the result found in most human perception studies, in which [pi] and [ki] are far more likely to be misheard as [ti] than the reverse. The directionality of errors in stop place identification will be discussed further in Chapter 5. The overall accuracy of stop place DT classification in the [i] context and indeed in all environments, the accuracy rate is lower than in the corresponding pairwise classifications (76.47%). As in the case of pairwise classification, the features employed in the context of [i] do not include formant onset features.

The task of three-way stop classification is more difficult, since the three types of stops overlap along most features. In many cases, a feature that reduces the overall entropy of all three stops at a node may not be a primary classifier in the pairwise classifications. This results in a lower overall accuracy for the three-way classifications.

5.2 DT classification of stop place in the [a] environment



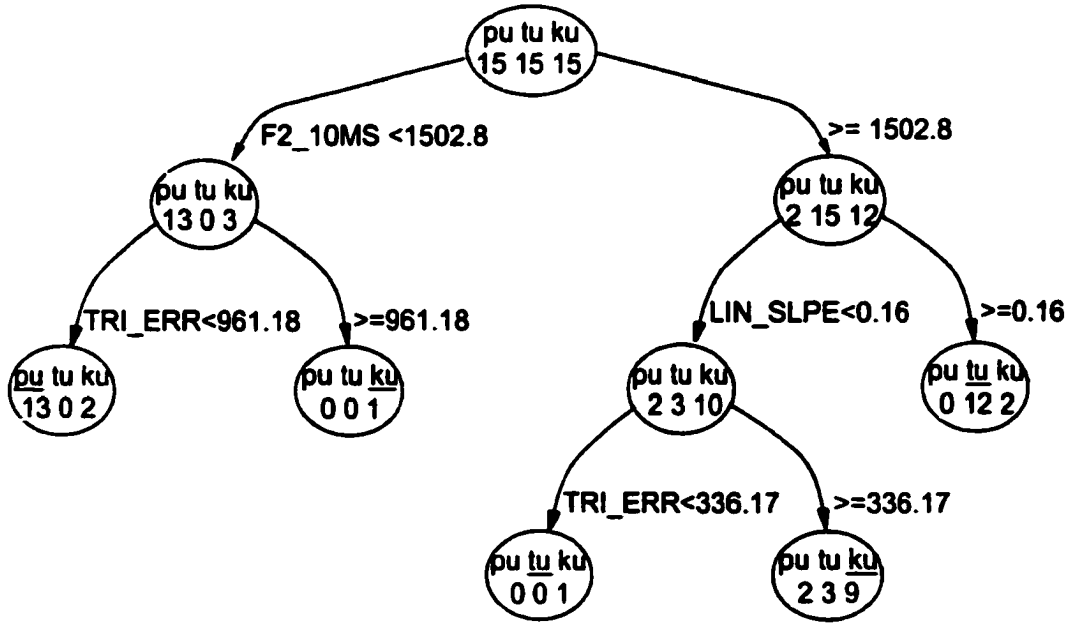
Accuracy: 79.17%

Figure 19. DT classification of test [pa], [ta], and [ka] tokens. The DT is trained on 171 CV tokens (57 [pa], 57 [ta], and 57 [ka]).

Using the two features F2 onset (F2_10MS) and the error of the triangular fit to the burst spectrum (TRI_ERR), the DT in Figure 19 classifies stops in the [a] context with 79.17% accuracy, which is only slightly higher than in the [i] context. CV tokens

with a low F2 onset are classified as *pa*. Tokens with a high F2 onset are further split by the feature TRI_ERR; a greater error indicates a [ka], and a lower error indicates a [ta]. These features work much the same way as in the multiple-feature DT classifications by CV pair (Section 4). The occurrence of F2 onset as a primary feature in the DT classification of stop place in C[a] position parallels its function as a primary cue in human perceptual studies (Cooper et al. 1952; Delattre et al. 1955; Kewley-Port 1982), particularly in the [a] vocalic context. In the set of test tokens shown in Figure 19, only one [pa] token is misclassified by its F2 onset. Misclassifications of [ta] and [ka] tokens occur more frequently.

5.3 DT classification of stop place in the [u] environment



Accuracy: 77.78 %

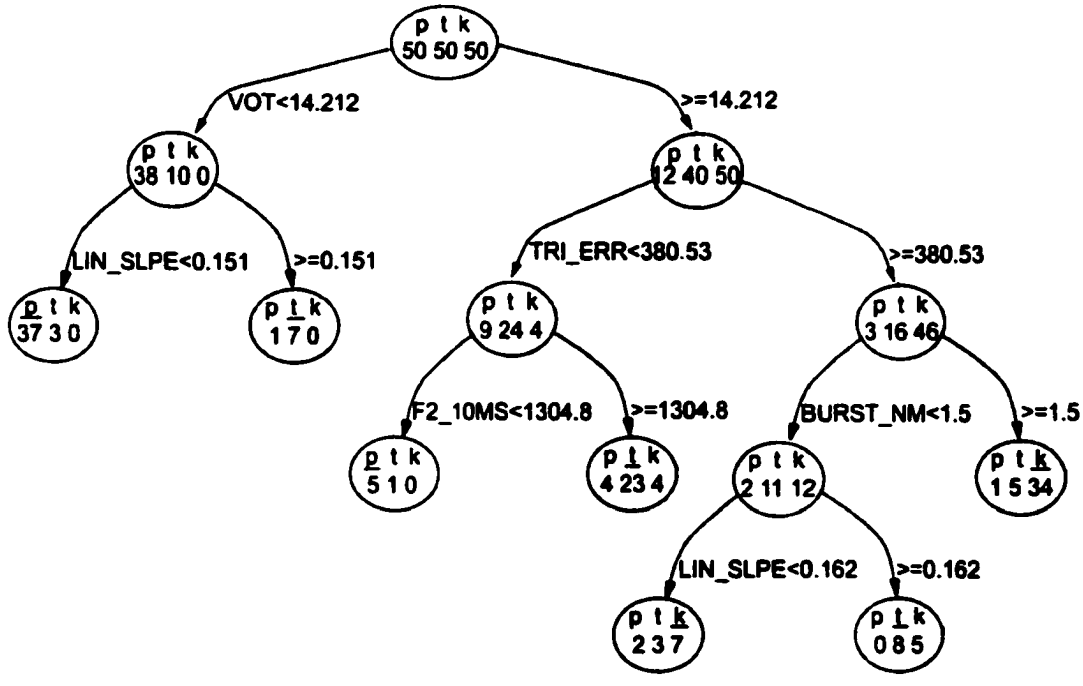
Figure 20. DT classification of test [pu], [tu], and [ku] tokens. The DT is trained on 171 CV tokens (57 [pu], 57 [tu], and 57 [ku]).

DT classification of stop place for C[u] tokens employs the features F2_10MS, LIN_SLPE, and TRI_ERR (Figure 20). As in the [a] context, the low onset of F2 primarily distinguishes bilabials from other stops. Stops with low F2 onsets are further classified by their TRI_ERR value, effectively separating [pu] tokens from [ku] tokens. Stops with high F2 onsets are further classified by LIN_SLPE, which separates the steep slope of most alveolar burst spectra from other stops. This classification of alveolars is not complete, however; a further branch separates alveolars from velars using the feature TRI_ERR. In the set of test data seen in Figure 20, TRI_ERR misclassifies 6 out of 15 tokens.

In the [u] context DT classifier, alveolars and velars are each associated with two separate leaf nodes, that is, two separate rules of classification. With the exception of TRI_ERR in the [tu] / [ku] classification, the acoustic features in this tree are also primary in the pairwise classifications. The accuracy rate is lower for this tree than for many pairwise classifications, due to the added difficulty of three-way stop classification.

5.4 DT classification of stop place in all environments

Finally, a multiple-feature DT classification of [p], [t], and [k] stops in all vowel environments most closely imitates the perceptual task that human listeners face in classic phoneme identification tasks involving CV stimuli where the vowel varies randomly (Figure 21).



Accuracy: 80.67%

Figure 21. Multiple-feature DT classification of stop place across all vocalic environments. The DT is trained on 516 CV tokens (172 [p]'s, 172 [t]'s, and 172 [k]'s).

The multiple-feature DT classification of stop place across all vocalic environments was trained on a set of 516 CV tokens and tested on a set of 150 test CV tokens, equally distributed across the nine possible CVs. The first major split is by the

feature VOT, which separates most labials and some alveolars from all other stops. Alveolars and bilabials in the short VOT branch are further separated by LIN_SLPE. The right side of the tree contains the stops with long VOTs, including all velars, most alveolars, and a few bilabials with uncharacteristically long VOTs. Upon closer inspection, I found that most of the bilabials in the long VOT branch were [pi] and [pu] tokens. The TRI_ERR test separates the velars (high TRI_ERR) from the remaining stops. The remaining alveolars and bilabials are distinguished by their F2 onset (F2_10MS). Bilabials that were successfully classified by their low F2 onset consisted of [pu] and [pa] tokens only.

The tokens in the branch on the far right of Figure 21 that were characterized by a long VOT and and a large TRI_ERR were mostly velars with some alveolars. The DT uses the feature BURST_NM to separate those velars with a multiple burst release, misclassifying a handful of [tu] and [ta] tokens in the process. The tokens with single stop bursts, consisting of about half alveolars and half velars, are distinguished by LIN_SLPE, which misclassifies [tu] tokens with uncharacteristically flat slopes and [ki] and [ku] tokens with uncharacteristically steep slopes.

The DT classification of all stops in all vocalic environments performs at a higher accuracy than many of the vowel-specific DT classifications, but at a lower accuracy than the majority of the pairwise DT classifications. Although the three-way task is more difficult for the tree than the pairwise DT classifications, the increase in the size of the data allows a much more robust training tree that is able to generalize to test data.

From the set of acoustic features extracted for each CV, test CV tokens of all three stop places in all three vocalic contexts can be characterized 80% of the time using

DT classification. This rate is comparable to human rates of identification in previous confusion studies. In the stop identification task by Winitz et al. (1972), for example, humans correctly identify [p], [t], and [k] at 80% accuracy when 100 msec of the following [i], [a], or [u] vowel was included in the stimuli. The set of features chosen to estimate discriminatory properties of the acoustic signal must correspond to some degree to cues of the signal the listeners rely on to identify stop place.

5.5 Summary and discussion

The DT classifications of stop place in each vowel context (Section 5.1-5.3) and across all vowel contexts (Section 5.4) share an interesting property. In each case, the bilabials are separated from the velars and alveolars in the top branch, usually by VOT or by F2_10MS. Only in the subbranches are the velars then separated from the alveolars, usually by a spectral feature, such as LIN_SLPE or TRI_ERR. In the DTs with more than three leaf nodes, the subbranching is usually dedicated to distinguishing alveolars from either bilabials or velars. The patterns found throughout the DT classifications in this section point to a hierarchy of similarity. Acoustic properties of voiceless unaspirated bilabials stops are the most distinct from other stop places, especially due to their short VOT and the low onset of their second formant. The acoustic signal generated from bilabial and velar stops are particularly distinct from one another across all vowel contexts. Voiceless unaspirated alveolar and velar stops, however, appear to occupy similar ranges of values along many acoustic properties, including VOT, F2 onset, and relative amplitude. Perhaps this hierarchy of acoustic similarity between [p], [t], and [k] accounts for the pattern found in languages with small stop inventories in which bilabial

stops contrast with a second stop with allophones [t] or [k] depending on the following vowel context (cf. Maddieson 1984).

Spectral properties are often the most reliable portion of the signal for distinguishing between the two stop places. Alveolar stops are likely to share acoustic properties with bilabial stops as well as velar stops, making their signal the least distinct of all three stop places, along the features studied here. In a sense, they occupy the middle ground, overlapping along features such as VOT with bilabials and properties such as F2 onset with velars.

6 Conclusion

In listener confusion studies, alveolars are often associated with the most confusions; bilabials and velars are more often confused for alveolars than any other stop place. The DT classifications in this chapter indicate that the confusions of bilabial and velar stop place for alveolars may be due in part to their similarities along many acoustic properties. The acoustic properties of bilabials and velars, on the other hand are much more distinct from one another, especially along features such as the second formant, VOT, and relative amplitude. The similarity does not explain, however, why listeners predominantly confuse bilabials and velars for alveolars but not the reverse. Properties of the acoustic signal responsible for the directionality of listener errors will be further discussed in Chapter 5.

Throughout this section, the DT classifications yielded lower accuracy rates for the [i] context as compared to the [a] and [u] contexts. This result mirrors many confusion studies that report higher listener errors for stops in the [i] context than stops in

other contexts. The DT classification results suggest that higher rates of listener errors in the context of [i] are due at least in part to an impoverished signal in this environment; there are less discriminatory cues for stop place available to the listener. Other human auditory and perceptual processing such as lexical bias or response bias certainly play a role in stop place identification as well, but the DT classifications show that there is a slight disadvantage to the [i] context based on the acoustic properties of the production of C[i] stops alone.

In the following chapter, I will present results from a perception study that investigates the relation between the featural and contextual salience estimated by DT classifications and the listener performance in a stop place identification task. As previously mentioned, the acoustic features extracted in this study most likely do not represent a complete set of acoustic cues to stop place. Nonetheless, listeners are expected to detect stop place more accurately than DTs, but their errors are expected to correspond to contexts and features poor in discriminant information as determined by the DT classifications in this section.

Chapter 4

Listener perception of stop place

1 Introduction

In the previous chapter, DT classifiers were used to estimate inherent featural and contextual informativeness for stop place. Listener identification of stop place is predicted to be more successful in contexts in which the acoustic signal is rich in discriminatory stop place information. Listener errors are expected to occur in contexts in which the acoustic signal is impoverished. The perception study presented in this chapter will evaluate these claims by examining the role of individual acoustic features in listener perception accuracy. CV tokens with canonical values for primary discriminatory acoustic features, as determined by DT classifications in Chapter 3, are expected to cause few listener errors in stop place identification. The perception experiment will address the following questions:

1. Are listeners more likely to make stop place confusions when a given context offers fewer cues?
2. If so, what are the acoustic features responsible for shifting the listeners' percept to another stop place category?
3. Which contexts and features exhibit asymmetrical errors?

A subset of the collected CV tokens described in Chapter 2 was selected as stimuli to test the role of individual features in listeners' responses. DT classifiers like those described in Chapter 3 were trained using the remaining CV tokens for the training set and the tokens chosen as stimuli for the test set. The significance of individual features in listener's responses and the relative rates of confusion by CV context in the perception study are then compared to the predictions of the DT classifiers to determine what part of listener errors is due to an ambiguous signal, and what part is due to other non-phonetic processing effects, such as lexical frequency. Factors that contribute to asymmetries in stop consonant errors are discussed in Chapter 5.

2 Method

When stop consonants are presented to subjects in high signal-to-noise conditions with the following vowel included, confusions are relatively rare (cf. Plauché 1997). Most perception studies interested in stop place identification errors attempt to increase confusion rates by truncating, masking, or filtering the stimuli. These methods of distortion are known to affect certain properties of the signal more than others (Bell et al. 1989). Since one purpose of this study is to investigate the relative roles of acoustic features, the usual methods of increasing confusion rates will not work. The CV stimuli in the current experiment were truncated 60 msec after vowel onset, which is sufficient for the identification of a consonant and the following vowel (Smits 2000 found that 60 msec corresponds roughly to the point at which stops can be identified by listeners at 100%). The CV stimuli were otherwise unaltered. Instead, listener errors were induced by

a cross-modal perception task, which induces errors by increasing the overall cognitive load of the subjects.

2.1 Stimuli preparation

The stimuli were chosen from a subset of the CV tokens previously described to test both the role of CV context and the role of individual acoustic features in listener identification of stop place. The DT classifiers of all three stops across all vocalic environments (Chapter 3, Section 4) are the DT models that closely match the conditions in the listener perception studies, in which the listener must identify the stop place of a CV stimuli from three choices. The estimations of contextual and featural salience from the DT classifications of CV pairs (Chapter 3, Sections 2 and 3), however, were chosen as the predictions against which to compare listener results because they showed the most variation by CV context.

For each pair of stops in each vocalic context (e.g., [pi] / [ti], [ti] / [ki], etc.), the most informative two to five features in the single-feature DT classifications of CV pairs (Chapter 2, Figures 6 to 8) were selected to be tested in the perception study (Table 1). The study was limited to the top two to five features for each CV pair because the selection process becomes increasingly complicated for each feature added. In cases where semi-redundant features were both found to be informative, such as F2_00MS and F2_10MS, only one feature of that class was chosen. Interacting features from multiple-feature DT classifications were also included.

CV Pair	Informative Acoustic Features
[pi] / [ti]	LIN_SLPE, VOT, F2_00MS
[ti] / [ki]	BURST_NM, TRI_ERR, NODE
[ki] / [pi]	REL_AMP, VOT, LIN_SLPE, BURST_NM, F2_00MS
[pa] / [ta]	LIN_SLPE, F2_10MS
[ta] / [ka]	LIN_SLPE*, NODE, TRI_ERR*
[ka] / [pa]	F2_10MS, VOT, NODE
[pu] / [tu]	F2_10MS, REL_AMP, LIN_SLPE
[tu] / [ku]	LIN_SLPE*, NODE, TRI_ERR*
[ku] / [pu]	TRI_ERR*, LIN_ERR, VOT*, REL_AMP

Table 1. Acoustic features tested for each CV pair. A '*' indicates features that interact with one another as determined by the multiple-feature DT classifications of the CV pair.

Let us take the [pa] / [ta] case as an illustrative example. According to the pairwise DT classifications in Chapter 3, [pa] and [ta] tokens are best distinguished by the features LIN_SLPE and F2_10MS (Table 1). To test the role of LIN_SLPE and F2_10MS in listener's responses, [pa] and [ta] tokens were chosen to represent both the canonical values along each feature and the overlap, or non-canonical values for each feature. The selected tokens were then presented to listeners as stimuli in the perception task, with the prediction that a [pa] token with overlap values along either LIN_SLPE or F2_10MS will be more likely to be confused for another stop place, such as [t], than a

[pa] with canonical values along both of those features. Let us look more closely at how these tokens are chosen.

The token selection process begins with histograms of the percentage of [pa] and [ta] tokens by their F2_10MS values (Figure 1) and their LIN_SLPE values (Figure 2).

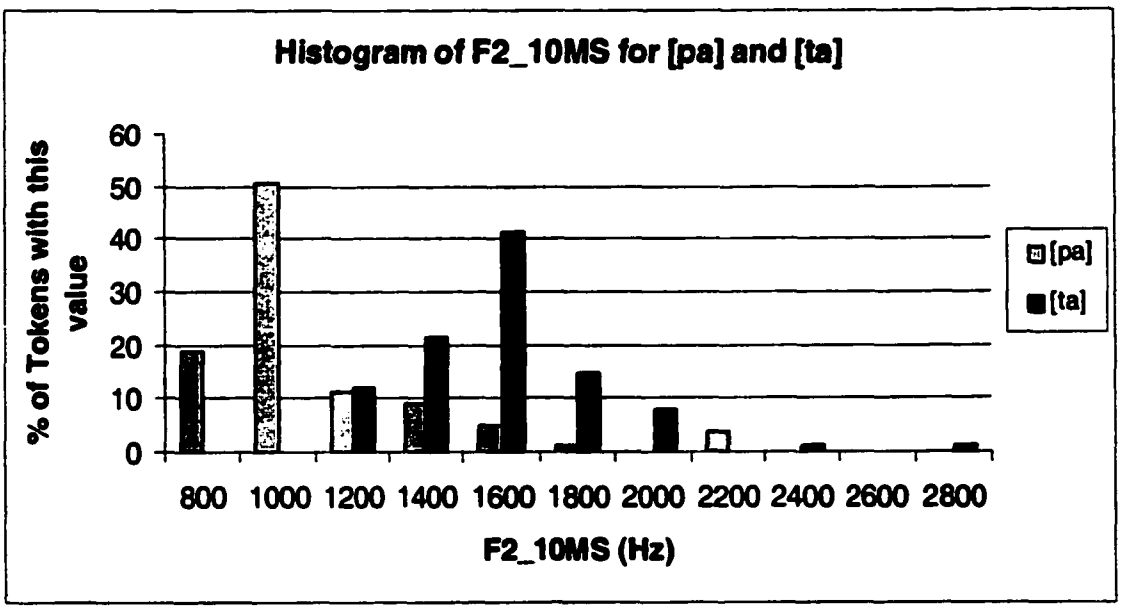


Figure 1. Histogram of [pa] and [ta] tokens by their F2_10MS values. The feature F2_10MS is binned with a bin size of 200 Hz. The [pa] peak is centered at 1100 Hz. The [ta] peak is centered at 1700 Hz. The crossover value is around 1300 Hz.

Most [pa] tokens have an F2 value at early vowel onset between 1000 and 1199 Hz (Figure 1), while [ta] tokens are centered around 1700 Hz. The crossover from values with a majority of [pa] tokens to values with a majority of [ta] tokens occurs around 1200 to 1399 Hz. The relatively low percentage of distribution overlap of [pa] and [ta] tokens causes the high accuracy rate in the DT classification of [pa] and [ta] along this feature. If listeners are aware of the discriminatory power of this feature, they may also rely on the

value of F2 at vowel onset in this study to discriminate between [p] and [t] in the context of [a]. CV stimuli were chosen to test this prediction.

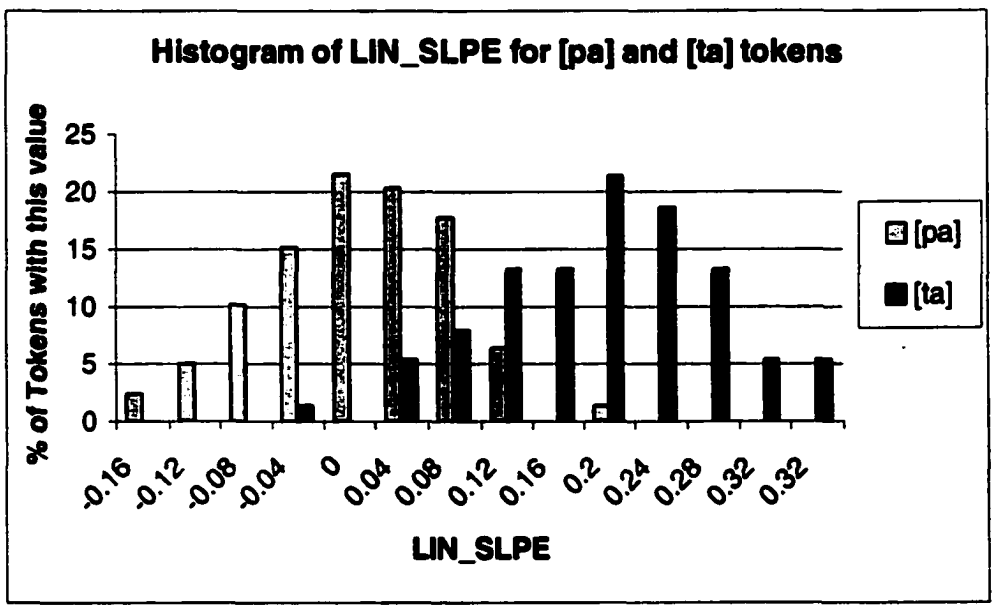


Figure 2. Histogram of [pa] and [ta] tokens by their LIN_SLPE values. The feature LIN_SLPE is binned with a bin size of 0.04. The [pa] peak is centered at 0.02. The [ta] peak is 0.22. The crossover value is around 0.1.

Most [pa] tokens have a very flat slope of the linear fit to the burst spectrum, that is, the LIN_SLPE value is around 0.02 (Figure 2). The majority of [ta] tokens have a steeper slope; their values for LIN_SLPE are centered around 0.22. The crossover from values possessed by a majority of [pa] tokens to values possessed by a majority of [ta] tokens occurs around 0.1. This feature also shows a relatively low percentage of distribution overlap between [pa] and [ta] tokens, which causes the DT to classify them with high accuracy along LIN_SLPE. If listeners are aware of the discriminatory power of this feature, they may also rely on the value of LIN_SLPE in this study to discriminate

between [p] and [t] in the context of [a]. CV stimuli were also chosen to test this prediction.

The first CV stimuli selected were the following *canonical* [pa] and [ta] tokens:

1. 2 canonical [pa] tokens – 2 [pa] tokens that lie within one standard deviation of the peak [pa] values along all tested features (1100 Hz for F2_10MS and 0.02 for LIN_SLPE).
2. 2 canonical [ta] tokens – 2 [ta] tokens that lie within one standard deviation of the peak [ta] values along all tested features (1700 Hz for F2_10MS and 0.22 for LIN_SLPE).

Non-canonical [pa] and [ta] tokens were selected from the *overlap* region of each feature. To test the role of the second formant in listeners' perception of stop place, non-canonical stimuli were chosen with respect to F2_10MS. The CV stimuli were also required to have LIN_SLPE values within one standard deviation of their peak values to control for the effect of LIN_SLPE. LIN_SLPE values for CV stimuli were required to lie within one standard deviation of 0.02 in the case of [pa] tokens and 0.22 in the case of [ta] tokens. The following *non-canonical* [pa] and [ta] tokens with respect to F2_10MS were selected:

3. 2 non-canonical [pa] tokens – 2 [pa] tokens that lie within one standard deviation of the F2_10MS crossover value (1300 Hz).
4. 2 non-canonical [ta] tokens – 2 [ta] tokens that lie within one standard deviation of the F2_10MS crossover value (1300 Hz).

To test the role of LIN_SLPE in listeners' identification of [pa] and [ta], the same selection process was performed with respect to LIN_SLPE. In order to control for the effect of F2_10MS, the CV stimuli were also required to have F2_10MS values within one standard deviation of their peak F2_10MS values. LIN_SLPE values for CV stimuli were required to lie within one standard deviation of 1100 Hz in the case of [pa] tokens and 1700 Hz in the case of [ta] tokens. The following *non-canonical* [pa] and [ta] tokens with respect to LIN_SLPE were selected:

5. 2 non-canonical [pa] tokens – 2 [pa] tokens that lie within one standard deviation of the LIN_SLPE crossover value (0.1).
6. 2 non-canonical [ta] tokens – 2 [ta] tokens that lie within one standard deviation of the LIN_SLPE crossover value (0.1).

Due to the relatively small set of data (70-80 tokens of each CV type), CV tokens that satisfied the exact peak and crossover criteria were not always available. In those cases, tokens were selected to conform to the criteria mentioned above as closely as possible. A total of 12 tokens was selected to test the role of the two features F2_10MS and LIN_SLPE in [pa] / [ta] confusions (Table 2). Values along all other features were checked for anomalies but otherwise were not controlled for.

Stimuli chosen	F2_10MS	LIN_SLPE
2 canonical [pa]'s	[pa] peak value	[pa] peak value
2 canonical [ta]'s	[ta] peak value	[ta] peak value
2 F2_10MS overlap [pa]'s	[pa]-[ta] crossover value	[pa] peak value
2 F2_10MS overlap [ta]'s	[pa]-[ta] crossover value	[ta] peak value
2 LIN_SLPE overlap [pa]'s	[pa] peak value	[pa]-[ta] crossover value
2 LIN_SLPE overlap [ta]'s	[ta] peak value	[pa]-[ta] crossover value

Table 2. Summary of the values along tested features of the 12 [pa] and [ta] stimuli. Tokens are selected to lie within one standard deviation of the values indicated.

For many CV pairs, three or four features were found to be informative by DT classification; the selection process is essentially the same as in the case of two features. Two canonical tokens and two overlap tokens are chosen for each CV condition for each of the three or four features. The overlap tokens must also lie within one standard deviation of the peak for all other features examined for that pair to control for the effect of the remaining features.

For some CV pairs, acoustic features that scored low in the single-feature DT classifications appeared as primary features in multiple-feature DT classifications, due to interacting effects among features. Though these acoustic features may not be useful when considered individually, in combination with other features they are apt at classifying CV pairs. For these cases, CV stimuli were selected from a scatterplot of the stimuli along two-dimensional *peak* zones and *crossover* zones.

The combinations of these algorithms resulted in the stimuli shown in Table 2 for the case of [pa] / [ta]. The number of CVs produced by each speaker was also carefully balanced. A total of 143 tokens was presented to subjects in the experiment (see Appendix D for the peak and overlap values of each stimulus).

2.2 Experimental design

127 native American English speakers aged 18 to 58 (90% were aged 18 to 22) participated in this perception study. The subjects were recruited from four different humanities classes. They had no formal phonetics training, had experience using computers, and reported no hearing loss. All subjects were paid for their participation in the study.

After receiving brief instructions, each subject provided personal information, such as age, initials, and dialect, and then began the experiment. The experiment was administered on a computer with headphones in a quiet room. During each experimental block, 25 to 30 images appeared in the center of the screen one by one at a fixed rate (one per second) (Figure 3). The images consisted of wingding characters that did not resemble orthographic letters of any language. The purpose of the cross-modal task was to force subjects to divert some of their cognitive load to the image task, causing greater errors in the phoneme identification task (Javkin, p.c.).

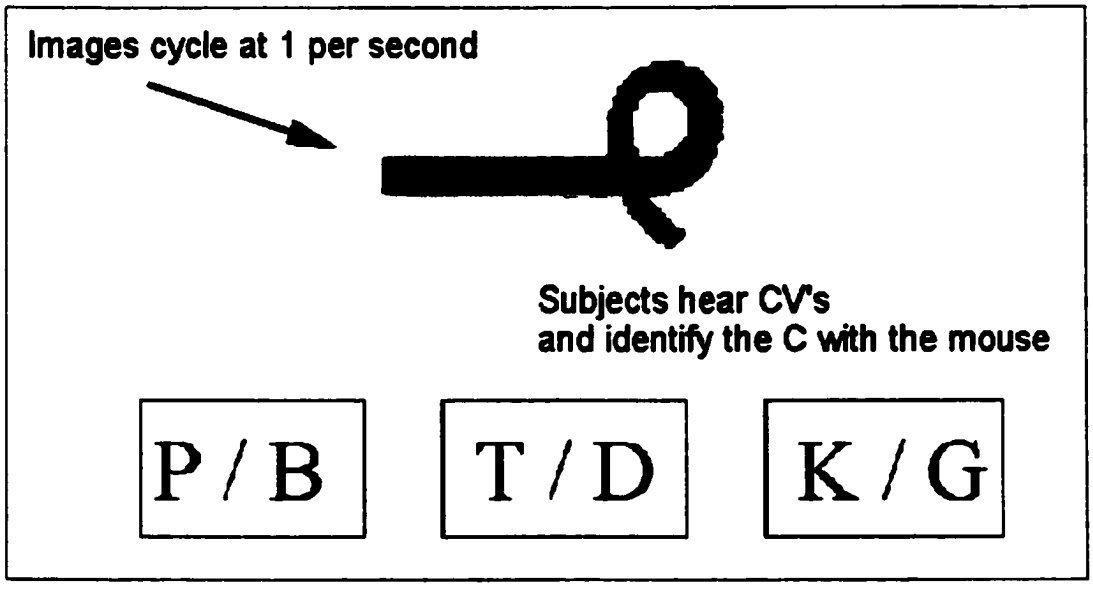


Figure 3. Computer interface during experimental blocks. Images appeared one by one in the middle of the screen. Three oversized stop category buttons remained at the bottom of the screen throughout each test block.

Subjects were instructed to pay attention to the images since they would be asked to identify whether or not they had occurred in the sequence, but they were not to click on the images during the task. Instead, they were instructed to use the computer mouse to click on one of the three consonant buttons at the bottom of the screen to identify the stop place of the CV stimuli they heard simultaneously over headphones. The rate at which the CV stimuli were presented depended on the response rate of the subject. Subjects were instructed to answer as quickly as possible, while still focusing on the cycling images in the center of the screen. If a subject took longer than three seconds to respond, no answer was recorded for the stimuli and the next stimuli was presented.

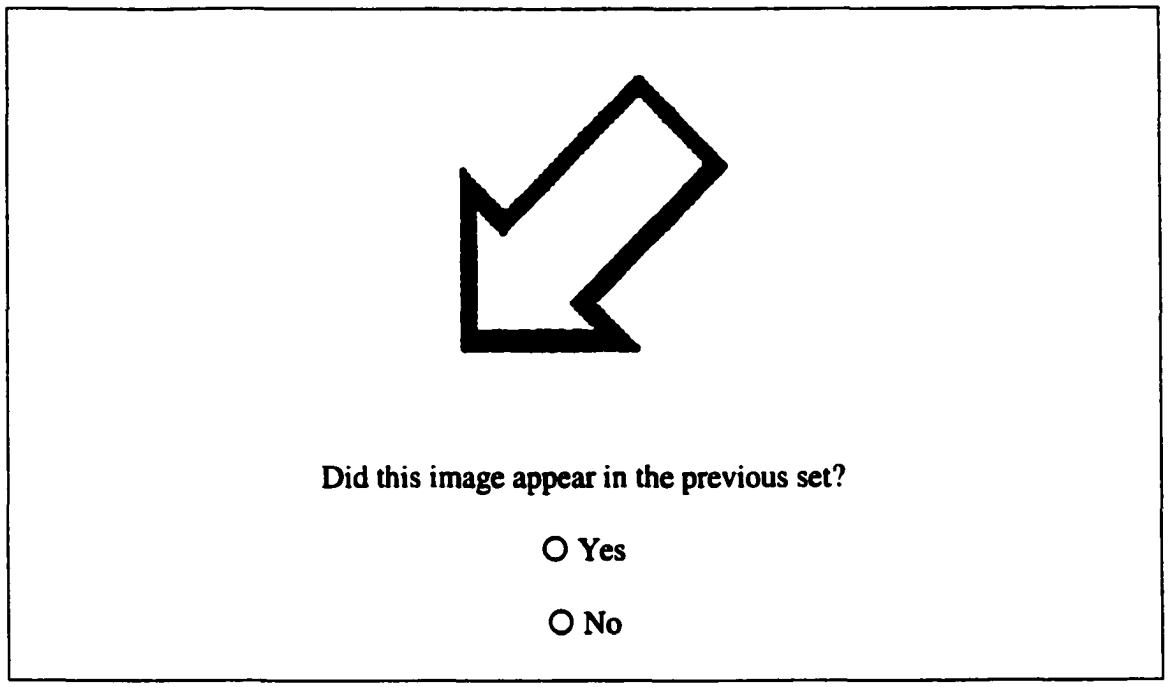


Figure 4. Computer interface at the end of an experimental block. Subjects answered the question by mouse-clicking in the appropriate circle.

After 23 or 24 CV stimuli, the images and audio stopped and a new screen (Figure 4) appeared on the computer. At the new screen, the subject was instructed to determine whether or not he or she had seen the indicated image in the previous set. This marked the end of an experimental block. Once the subject selected *yes* or *no*, a new experimental block began (Figure 3).

The experimental interface was designed so that the mouse was free to be used entirely for the audio task during the cross-modal portion of the test block. The buttons at the bottom of the screen were large and turned grey when the mouse pointer crossed them. This allowed the subject to identify consonants using mainly peripheral vision, focusing instead on the images during the test block.

The entire experiment consisted of a short training task, two test blocks, a short break (one to two minutes), two test blocks, another short break, and finally, two more test blocks. The entire experiment took about 20 minutes to complete, including instruction. The breaks were designed to allow subjects to rest briefly between blocks, since the image portion of the task was fatiguing.

Two versions of the experiment were conducted, each using approximately half of the subjects. In one version of the test (Version A), CV stimuli with different vowels were presented randomly in all six blocks. In another version (Version B), the experimental blocks contained CV stimuli with identical vowels: the training block contained C[i] stimuli, the first two blocks were C[a] stimuli, the middle two blocks contained all C[i] stimuli, and finally, the last two blocks contained all C[u] stimuli. To run a maximum number of subjects, the experiment was also conducted on two separate computers at once; one computer ran the experiment notably faster than the other.

Eight additional subjects volunteered to participate in a version of the study in which only the image portion of the task was tested. These subjects heard the stimuli but were instructed to respond *p/b* for all stimuli and to instead focus all of their attention on the image recall task. Their results provided a baseline for the image portion of the task against which all other subjects could be compared.

3 Results

3.1 Overall results

Of the 127 total speakers, none performed below chance on the image recall task, though four performed exactly at chance. A series of one-way ANOVAs were first conducted to determine whether response rates to CV stimuli were significantly affected by the listener, the age of the listener, the version of the experiment, or the machine used. Overall responses were significantly affected by the machine used ($F(1, 18159) = 21.8$), age of the listener ($F(15, 18145) = 2.535$), listener ($F(126,18034) = 1.443$), but not by the experimental version ($F(1, 18159) = 0.001$).

Since almost 90% of the listeners (110 out of 127) were between the ages of 18 and 22, a one-way ANOVA on the effects of age, listener, machine used, and experimental version was conducted on this subset only. The results showed that age ($F(4, 15725) = 0.19$), dialect ($F(2,15727) = 1.051$), listener ($F(109, 15620) = 1.203$), and version of the test ($F(1, 15728) = 0.016$) were not significant factors in listeners' responses. Whether the subjects used the fast or slow machine did show significant effects for these subjects ($F(1, 15728) = 15.692$). Since this subgroup of subject responses shows no significant effect by age and by listener, only results from the 18 to 22 age group were used for further analysis. Not surprisingly, for this group as well as for all results, the overall listener responses were also affected by the following vowel ($F(2, 15727) = 18.702$), and the consonant of the stimuli ($F(2, 15727) = 16246.999$).

The overall confusion matrix by CV is shown in Table 3. Velars caused the most problems for listeners in all three vocalic contexts, in which they are often confused for

alveolars. Alveolars and labials have higher rates of correct identification overall but their rates show more variation by the following vowel. Alveolar stops are more likely to be confused for a bilabial place of articulation when the following vowel is an [i]. Confusions for velar place are more common when the following vowel is an [a]. [tu], on the other hand, is rarely confused for another place of articulation. In the case of bilabial stops, [pi] is most often confused for [ti] but [pu] is equally likely to be confused for an alveolar or velar stop place as it is to cause a *no answer* (NA) response. [pa] is rarely confused by listeners in this study. With the exception of [pu], NA responses remain below 3% for all CVs.

Stimuli /	[p]			[t]			[k]		
Response	[i]	[a]	[u]	[i]	[a]	[u]	[i]	[a]	[u]
P/B	**93.2**	**94.7**	**87.3**	4.5	*2.8*	*0.5*	3.6	*0.3*	*0.6*
T/D	3.4	*1.0*	4.3	**90.8**	**87.4**	**96.2**	15.0	13.5	10.8
K/G	*0.7*	*1.7*	5.2	*2.0*	7.6	*1.1*	**79.4**	**84.2**	**86.1**
NA	*2.7*	*2.5*	3.2	*2.7*	*2.2*	*2.3*	*1.9*	*2.0*	*2.5*

Table 3. Confusion matrix. Response rates are given in rounded percentages. *NA* indicates a *no answer*, in which the subject took longer than three seconds to make a decision. Correct responses are shown in bold. Rare responses are italicized.

The correct response rates for each CV context were separated by canonical and non-canonical stimuli results (Figure 5). For each CV context, four stimuli were canonical, with the exception of [pu], which contains only three canonical stimuli. The

number of non-canonical stimuli by CV ranges from 12 to 20, depending on the number of relevant acoustic features.

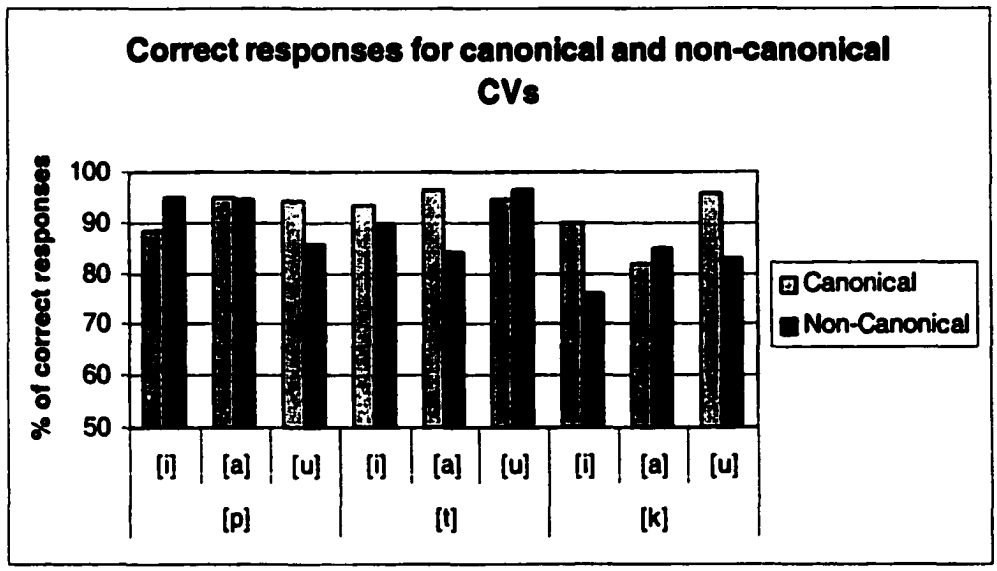
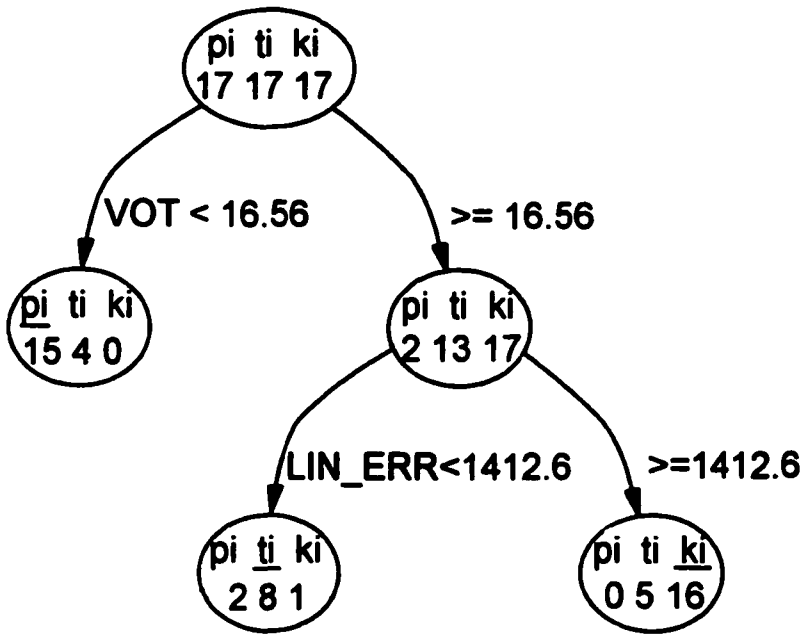


Figure 5. Correct responses by CV for canonical and non-canonical stimuli.

For the majority of CV conditions, the non-canonical tokens caused more listener errors than the corresponding canonical tokens. Exceptions to this generalization include [pi], [pa], [tu], and [ka], for which the non-canonical tokens were identified at the same or slightly higher rates. Both canonical and non-canonical [ka] tokens are often confused for alveolars. [pu], [ta], [ki], and [ku] show the largest drop in identification accuracy in the non-canonical condition, suggesting that for these four CV contexts, listener perception is affected by a stimuli's value along relevant acoustic features. The exact features responsible for the increased confusion rates will be examined in Section 3.3.

3.2 Contextual salience

In the previous chapter, multiple-feature DT classifications of stop place ranked the relative informativeness of vocalic contexts by their accuracy rates: The [a] context is more informative than the [u] context, which is in turn more informative than the [i] context (Chapter 3, Section 4). For a given vocalic context, the number of correctly classified tokens can be calculated for each stop. For example, in the multiple-feature DT classification of stop place in the [i] context, 15 out of 17 [pi] tokens, 8 out of 17 [ti] tokens, and 16 out of 17 [ki] tokens were correctly classified (Figure 6). The accuracy rates of classification by CV can be compared with human response rates by CV to test whether listener's performance is subject to differences in the inherent informativeness of a given context.



Accuracy: 76.47%

Figure 6. DT classification of unseen [pi], [ti], and [ki] tokens. All CV tokens in a given leaf node are classified by the underlined CV of that node. The DT was trained on 165 CV tokens (55 [pi], 55 [ti], and 55 [ki]).

The test sets for each of the original DT classifiers were relatively small (15 to 17 tokens per CV). As a result, the DTs and their accuracies were sensitive to the individual tokens in the test set. To normalize for any differences caused by individual tokens, the multiple-feature DT classifications by CV pair were rerun such that the DT test set was composed of the same set of CVs that were presented as stimuli to the human subjects. The *adjusted* DTs are similar in structure to the original DTs, but accuracies and thresholds vary slightly (Table 4).

	Original DT			Adjusted DT		
	Training	**Test**	**Accuracy**	**Training**	**Test**	**Accuracy**
[pi]	55	17	**88.2**	59	13	**84.62**
[ti]	55	17	**47**	59	13	**53.85**
[ki]	55	17	**94**	59	13	**92.3**
[pa]	57	16	**93.75**	58	14	**100**
[ta]	57	16	**75**	58	14	**57.14**
[ka]	57	16	**68.75**	58	14	**71.43**
[pu]	57	15	**86.67**	56	16	**81.25**
[tu]	57	15	**80**	56	16	**81.25**
[ku]	57	15	**66.67**	56	16	**68.75**

Table 4. Original and adjusted multiple-feature DT classification scores. The training and test sets for the original DTs were randomly selected. The test set of the adjusted DTs are the same as the set of experimental stimuli. The training set of the adjusted DTs is composed of the remaining CVs.

To determine the extent to which listener accuracy was affected by the inherent informativeness of a given context, the percentage of correct listener responses for a given CV was plotted by the adjusted DT classification accuracy by CV, as presented in Table 4. Listener response rates did not correlate significantly with the adjusted DT accuracy (Figure 7).

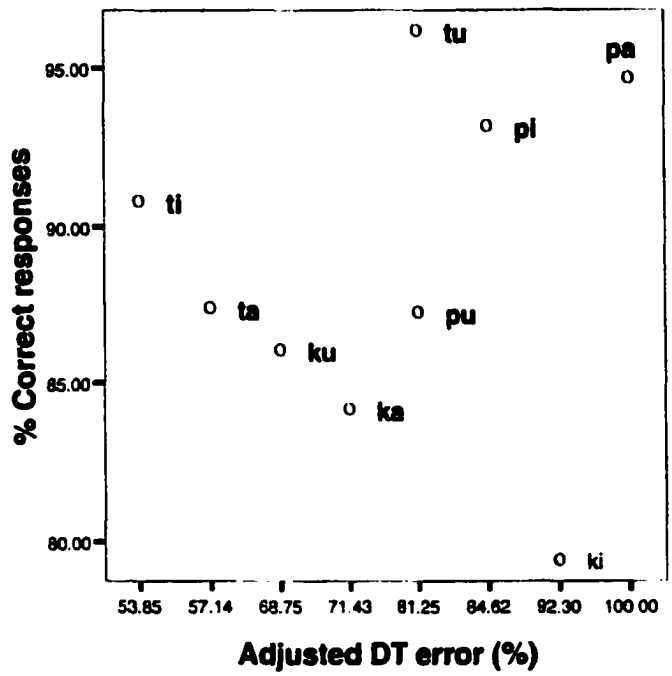


Figure 7. Human versus DT accuracies by CV. The x-axis represents the adjusted DT accuracies found in Table 4. The y-axis represents the percentage of correct responses by CV in the perceptual study.

Both listeners and DTs correctly identified [pa], [pi], and [tu] tokens, but all other CV tokens failed to correlate. This discrepancy may be due to aspects of the DT training algorithm, which is based on reducing the entropy at each node. The resulting tree structure may perform better for some CVs than others, depending solely on the position of the node. Also, although the three-way DT classifications are the DT tasks that most closely resemble the perceptual experiment, they may not be the most accurate representation of the information available in the signal for a given CV. For example, the DTs start with equal priors for all CV contexts. Indeed, in the perceptual experiment,

subjects hear equal numbers of each CV context, but the frequency of a CV context in the lexicon is a factor in listeners' responses, as is further discussed in Chapter 5, Section 3.

Let us turn now to DT classifications by CV pair. The accuracy rates of a DT for a given pair of CVs ([pi] / [ti], for example) can be compared with listeners' accuracy in distinguishing between the same two CVs. The best way to approach this is to investigate percentage error of both DTs and listeners. The percentage error in a multiple-feature DT classification (Chapter 3, Section 3) is simply the complement of the accuracy (= 100% – accuracy). Listener error between [pi] and [ti] tokens can be calculated by averaging the percentage of [pi] stimuli misheard as [ti]'s and the percentage of [ti] stimuli misheard as [pi]'s.

But first, because the test sets for the DT classifications were relatively small (15 to18 tokens per CV), the fits, and therefore the accuracies, are highly sensitive to the individual tokens in each test set. The multiple-feature DT classifications by CV pair were rerun such that the DT test set was composed of the same set of CV stimuli presented to the human subjects. The adjusted DTs are similar in structure to the original DTs, but accuracies and thresholds vary slightly (Table 5).

	Original DT			Adjusted DT		
	Training	**Test**	**Accuracy**	**Training**	**Test**	**Accuracy**
[pi] / [ti]	55x2	17x2	**85.29**	59x2	13x2	**76.92**
[ti] / [ki]	55x2	17x2	**79.41**	59x2	13x2	**88.46**
[ki] / [pi]	60x2	18x2	**91.67**	62x2	15x2	**100**
[pa] / [ta]	57x2	18x2	**94.44**	60x2	14x2	**100**
[ta] / [ka]	57x2	16x2	**84.38**	58x2	15x2	**66.67**
[ka] / [pa]	60x2	16x2	**100**	58x2	14x2	**100**
[pu] / [tu]	60x2	15x2	**96.67**	56x2	17x2	**97.06**
[tu] / [ku]	57x2	15x2	**76.67**	56x2	16x2	**71.88**
[ku] / [pu]	57x2	16x2	**93.75**	57x2	16x2	**78.12**

Table 5. Original and adjusted multiple-feature DT classification scores. The training and test sets for the original DTs were randomly selected. The test set of the adjusted DTs are the same as the experimental stimuli. The training set of the adjusted DTs is composed of the remaining CVs.

The adjusted DT classification errors were plotted by the percentage of listener errors by CV pair (Figure 8). Although the correlation between listener and adjusted DT error is weak ($R^2 = 0.53$), there does appear to be a set of CV pairs that both the DT and listeners identified accurately more than 95% of the time ([ki] / [pi], [pu] / [tu], [pa] / [ta], and [ka] / [pa]) as well as a set of CV pairs with low DT accuracies for which listener error rates were elevated ([ti] / [ki], [ta] / [ka], [pi] / [ti], [tu] / [ku]). The pair [ku] / [pu] had low accuracy for DT classification, but caused little error for listeners.

As predicted, listeners had more trouble discriminating between CV pairs for which less discriminatory stop place information was available, as estimated by the

multiple-feature DT classifications. In particular, according to DT classification accuracies, alveolars and velars were similar along most acoustic features in all three vocalic contexts. Likewise, listeners appeared to have more trouble distinguishing between these two stop places than any other CV pair in the perception study.

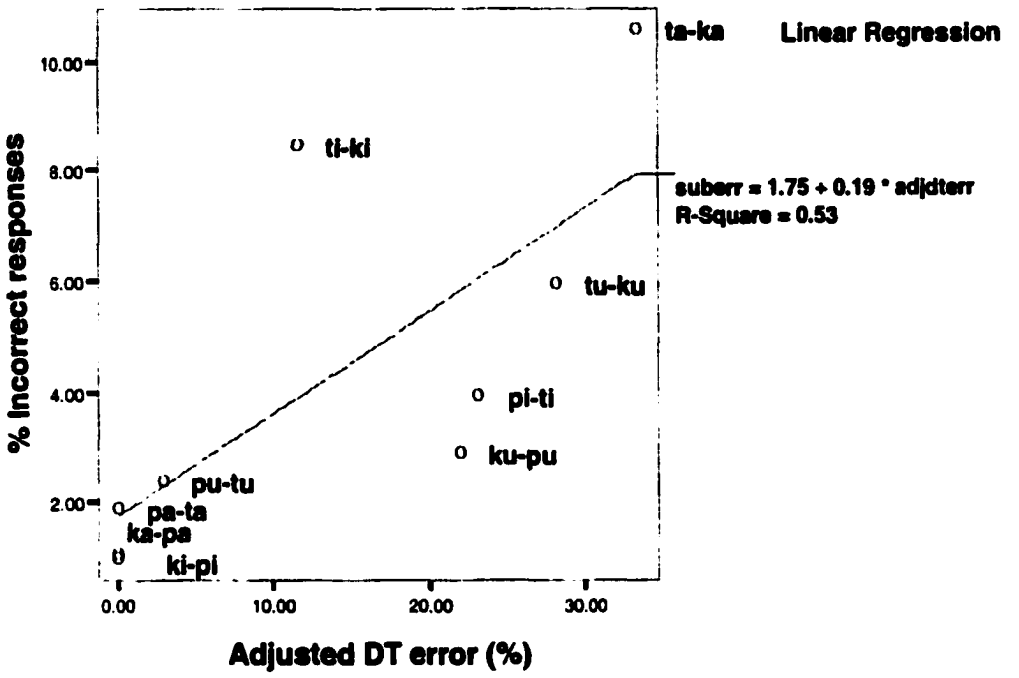


Figure 8. Linear regression of percentage of listener error by the adjusted DT error.

3.3 Featural salience

The overall results (Section 3.1) showed that listener error was greater for non-canonical [pu], [ta], [ki], and [ku] tokens than for their canonical counterparts. For these CVs, listener perception appears to be affected by the CV stimuli value along the tested

acoustic features. In this section, I determine which individual acoustic features are used by listeners in the identification of stop place for all CV tokens in turn.

The role of individual features in listener error rates is determined by a linear regression model of the listener error for a given CV stimuli by its values along all acoustic features. When a linear regression of the percentage of correct responses per stimuli by the feature value for each stimulus is found to be significant, it indicates one or more features that were used by listeners to identify tokens of that CV context. Although monotonic fits to the data are expected, the distribution of the data need not be linear to demonstrate the significance of a feature in listener error rates.

[pi] identification

Listeners had very little difficulty identifying the stop place of [pi] stimuli (93.2% overall), even when the stimuli were non-canonical along relevant acoustic features (95.5%). The stop place confusions that did occur were caused by four stimuli in particular (Figure 9).

Listener performance for [pi] tokens varied significantly by REL_AMP (Beta = 0.564, $p < 0.005$). Bilabials have the lowest relative amplitude values of all three stop places in the [i] context (peak = 0.1), but a scatterplot of listener identification rates of [pi] stimuli by REL_AMP values shows that the closer the [pi] stimulus is to the [pi] mean REL_AMP value (0.1), the more likely it is to be correctly identified. Stimuli with lower REL_AMP values were more likely to be confused for other stop places. This result is unexpected, since both alveolars and velars tend to have higher REL_AMP values. It is likely that this result is due to listeners having more trouble identifying stop

place when the stop burst is quiet. Additionally, the low relative amplitude of the burst may indirectly affect listener response rate by impeding other important cues to stop place, such as gross spectral properties of the burst.

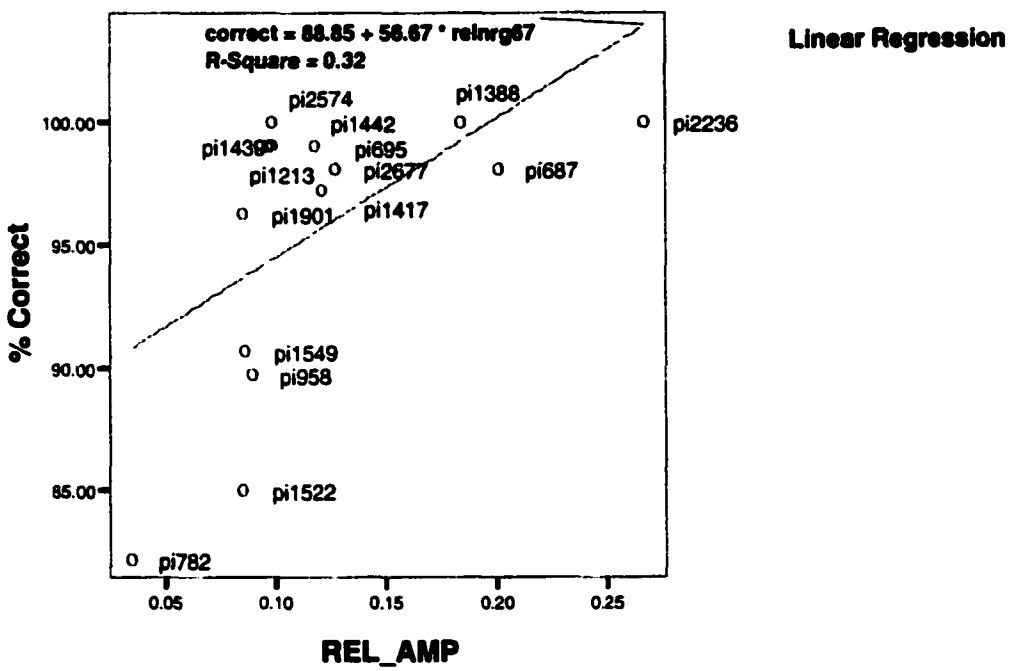


Figure 9. Linear regression of percentage of correctly identified [pi] stimuli by REL_AMP.

[pa] identification

Listeners made very few errors identifying [pa] stimuli in this experiment (5.2% of all responses), whether the stimuli were canonical or non-canonical. The difference in the percentage of correct responses by individual stimuli significantly varies by TRI_ERR (Beta = 0.576, $p < 0.005$). The scatterplot shows that [pa] tokens with values of TRI_ERR closest to the mean (0) caused listener confusions (Figure 10). This result was

unexpected since TRI_ERR was not predicted to be a primary feature by DT classifications.

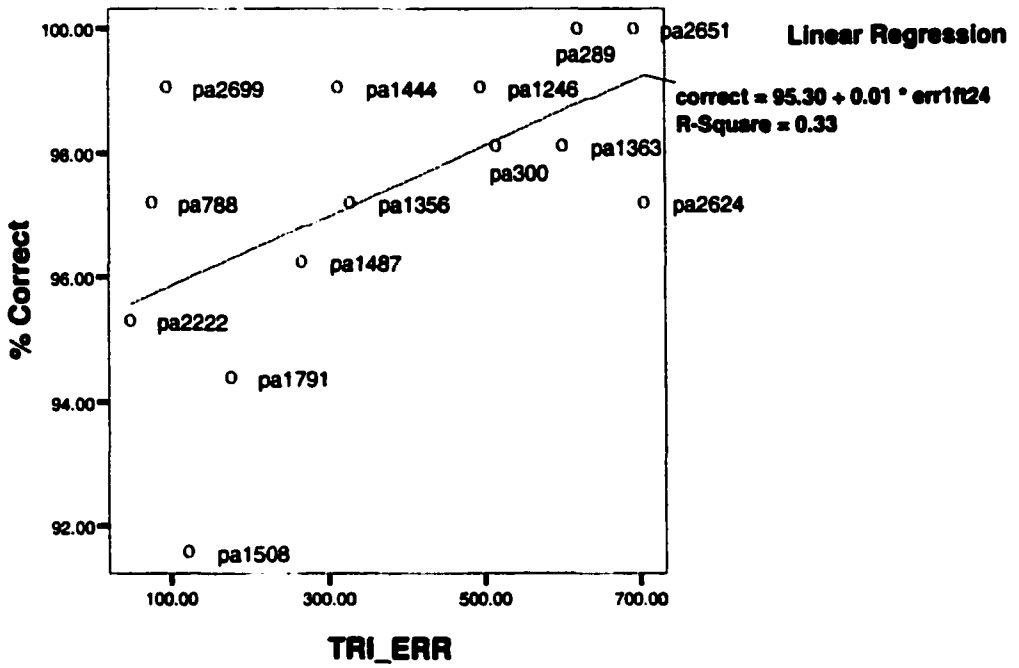


Figure 10. Linear regression of percentage of correctly identified [pa] tokens by TRI_ERR.

[pu] identification

Listener identification rates were significantly lower for non-canonical [pu] stimuli (85.8%) than for canonical stimuli (94.2%). The [pu] context was predicted to have the following primary features for listener detection of stop place: F2_10MS, REL_AMP, LIN_SLPE, TRI_ERR, LIN_ERR, VOT (as compiled from the [pu] / [tu] classifications and the [pu] / [ku] classifications). A stepwise linear regression of all acoustic features showed that F2_10MS and LIN_SLPE contributed significantly to the percentage of

correct responses (Beta = -0.506, 0.448, respectively; $p < 0.005$). No other features were found to significantly affect listeners' response rate.

The scatterplot of the stepwise linear regression (Figure 11) shows that listeners made the fewest errors in [pu] identification for [pu] tokens with an F2 onset lower than1500 Hz and a LIN_SLPE between 0.05 and 1.0. These feature ranges correspond to the distributional mean of [pu] tokens, ranges in which 80% or more of the [pu] tokens lie. Both of these features were predicted by the DTs to be useful in [pu] discrimination. Listeners did not rely on REL_AMP, TRI_ERR, VOT, LIN_ERR, or other acoustic features for the identification of [pu] tokens.

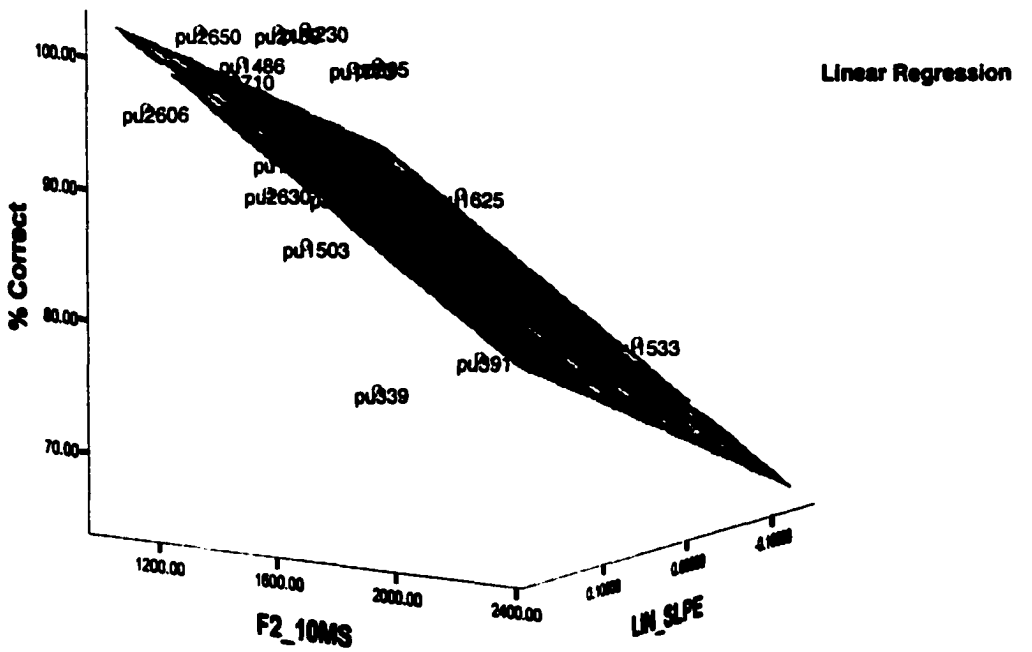


Figure 11. Stepwise linear regression of correct responses to [pu] stimuli by their values along LIN_SLPE and F2_10MS ($R^2 = 0.5$).

[ti] identification

Listeners rarely misidentified the stop place of [ti] stimuli, regardless of whether the stimuli were canonical (93.4%) or non-canonical (89.6%). A multiple linear regression model indicated no significant acoustic feature effects in the identification rate of [ti] stimuli. Scatterplots of listener response rate by individual feature values confirmed this result.

[ta] identification

Listener identification rates were significantly lower for non-canonical [ta] stimuli (84.1%) than for canonical stimuli (96.4%). According to the DTs, the following features were most likely to be used by listeners for the discrimination of [ta] from other stop places: F2_10MS, VOT, LIN_SLPE*, NODE, TRI_ERR*. All acoustic features were input to a stepwise linear regression model to investigate their effects on listener responses. The linear regression model failed for all features. Upon closer examination, I found that one token in particular (*ta480*) had triggered a large percentage of listener error but was anomalous along many feature values with respect to other [ta] tokens. Once this particular stimulus was removed, the linear regression revealed significant effects for the features VOT and LIN_SLPE (Beta = 0.711, -0.458, respectively, $p < 0.005$). A [ta] token with a VOT between 5 and 17 msec and a LIN_SLPE greater than 0.12 was correctly identified more often than tokens with longer VOTs or flatter spectral slopes. The ranges of VOT and LIN_SLPE correspond to peaks in the distribution of all [ta] values along the two features. Both VOT and LIN_SLPE are rich in stop place information in the [ta] context.

[tu] identification

Listeners had very little trouble identifying [tu] stimuli, both in canonical (94.5%) and non-canonical conditions (96.6%). A linear regression on all relevant features revealed that the value of VOT significantly affected listener's responses (Beta = -0.468, $p < 0.005$). A scatterplot shows that listener errors, though relatively few, become slightly more likely as the VOT increases (Figure 12). Dynamic cues not captured by the extracted acoustic features, such as the steep drop in F2 at the onset of the vowel, are also likely to have affected listener responses.

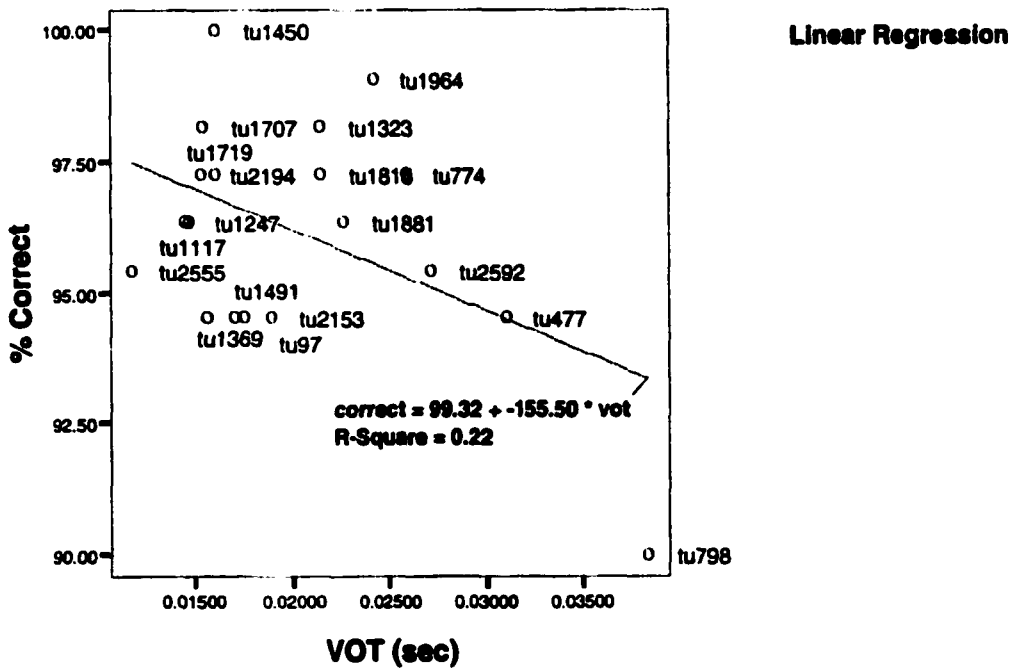


Figure 12. Linear regression of the percentage of correctly identified [tu] tokens by VOT.

[ki] identification

Listener identification rates were significantly lower for non-canonical [ki] stimuli (75.9%) than for canonical stimuli (90%). DT classifications of [ki] tokens revealed that the following features of the acoustic signal provided the most discriminatory power: BURST_NM, TRI_ERR, NODE, REL_AMP, VOT, LIN_SLPE, F2_00MS. A stepwise linear regression model of percentage of listener error by all acoustic feature values revealed no significant feature effects. Scatterplots showed that one [ki] token was anomalous along most features (Figure 13), especially along VOT, an otherwise correlating feature. When the multiple regression model was rerun excluding this stimulus (*ki2148*), VOT was found to significantly correlate (Beta = 0.558, $p < 0.005$) to listener response rates. With the exception of the one anomalous token, [ki] stimuli with a VOT shorter than 24 msec were more likely to cause perceptual errors than those with VOTs between 24 and 50 msec. This range corresponds to the range of VOT values for the majority of all [ki] tokens in the database.

The feature F2_10MS correlated at just under significance for all [ki] tokens, excluding the anomalous one ($p < 0.005$). A scatterplot showed that all of the [ki] tokens that caused listener errors 30% of the time or more had F2 onsets below 2300 Hz. This result was somewhat unexpected, since F2_10MS was not one of the primary acoustic features in DT classifications. A related feature, F2_00MS was primary in DT classifications but was not found to significantly affect listener response rate.

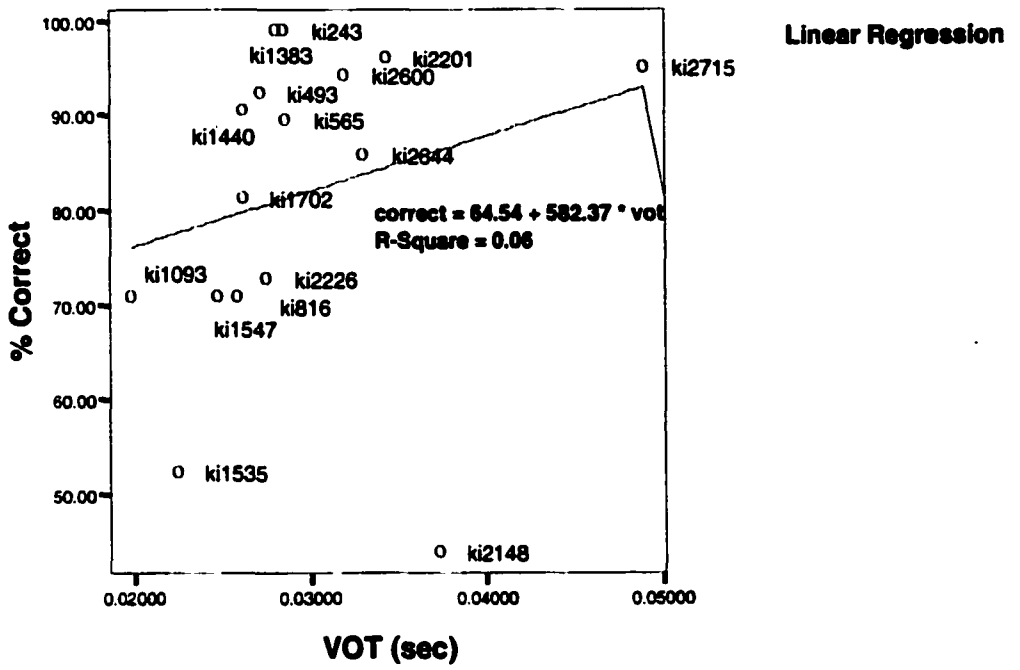


Figure 13. Linear regression of percentage of correct listener responses for each [ki] stimuli by VOT values. When the token *ki2148* was excluded, the R^2 value increased to 0.31.

[ka] identification

The [ka] stimuli were misclassified by listeners 15% of the time or more, regardless of whether the stimulus was canonical or non-canonical along the relevant features. There are two explanations for the pattern of [ka] tokens: either an untested feature that was not controlled for caused listener errors, or the [ka] context is particularly prone to causing listener error, whether the individual tokens have characteristic values along relevant features or not. We know from previous phonetic studies that velar stop place is more likely to cause listener confusions than other places, but this trend is not particular to the [a] context.

The stimuli were most likely not correctly selected to test the relevant features. Indeed, when we examine the listener rates by [ka] stimuli, we see that one token *ka284*, chosen to be canonical, was misidentified by listeners more often than any other [ka] stimulus. The token's values along all possible features (LIN_SLPE, NODE, TRI_ERR, F2_10MS, VOT, as well as LIN_ERR, F2_00MS, BURST_NM, and REL_AMP) are less than one standard deviation from the peak value for [ka] tokens. A scatterplot and linear regression of percentage of correctly identified [ka] stimuli by their value along VOT, however, indicates that VOT is a significant factor (Beta = 0.664, $p < 0.005$) in listeners' performance (Figure 14). More importantly, it shows that the token *ka284* is one of only two stimuli to have a VOT value shorter than 15 msec. Although this value is only one standard deviation from the peak VOT value of all [ka] tokens (17.5 msec), it is clearly short enough to cause listener confusion. Multiple linear regressions revealed no other significant features in listener performance.

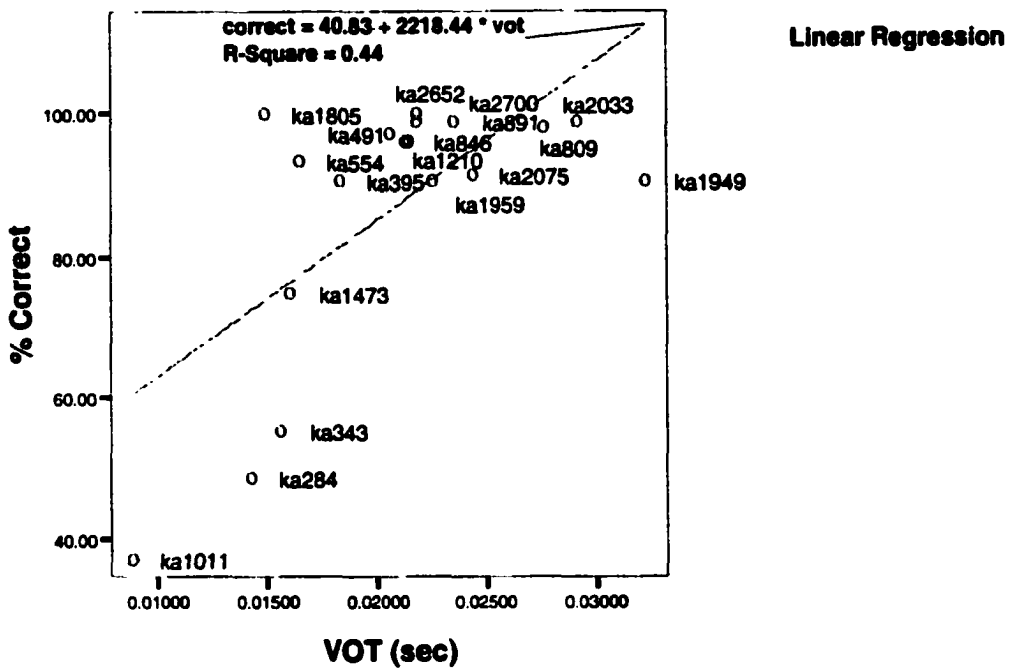


Figure 14. Linear regression of percentage of correct responses for each [ka] stimuli by VOT value.

[ku] identification

Listener identification rates were significantly lower for non-canonical [ku] stimuli (83%) than for canonical stimuli (95.7%). DT classifications of [ku] tokens relied primarily on the following acoustic features: LIN_SLPE, NODE, TRI_ERR, LIN_ERR, VOT, REL_AMP, and F2_00MS. None of these features revealed significant effects on listener response rates for [ku] tokens. NODE correlated at near significance (Beta = -0.47, $p < 0.005$), however. Those [ku] tokens with NODE values between 1 and 3 KHz were rarely confused for other stop places. Those with higher NODE values were much more likely to cause listener error.

Scatterplots of listener response rates by the remaining feature values did reveal one anomalous [ku] token (*ku1511*) (Figure 15). When this token was removed, F2_00MS was found to significantly affect listener response rate (Beta = -0.724, $p < 0.005$). [ku] tokens with F2 onsets at frequencies higher than 1750 Hz were more likely to be confused by listeners for other stop places than [ku] tokens with F2 onsets between 1400 Hz and 1750 Hz. VOT correlated at near significant rates (Beta = 0.336, $p < 0.005$). Though the correlation was much weaker, a scatterplot showed that all [ku] tokens correctly identified at 80% accuracy or higher had VOTs longer than 18 msec, corresponding exactly to the range of VOT values for the majority of all [ku] tokens in the database.

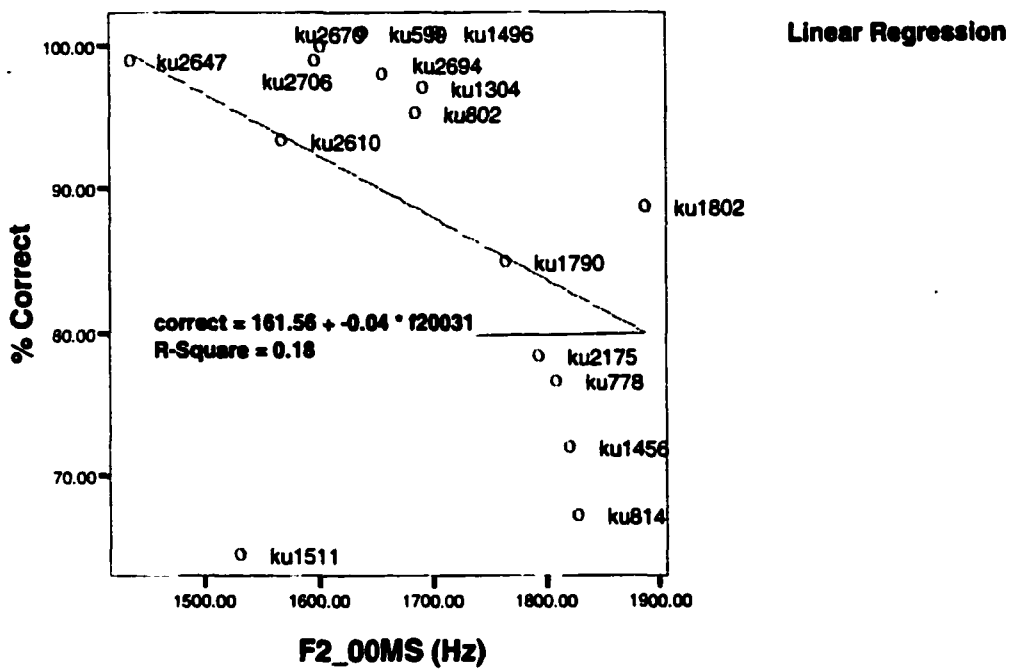


Figure 15. Linear regression of percentage of correct listener responses for each CV by the value of F2_00MS for that CV. When the token *ku1511* was removed, the R^2 value increased to 0.52.

The percentage of correct responses per [ku] token appears to decrease exponentially as the F2 at voicing onset approaches 1900 Hz. There is not enough data to make claims about the shape of the fit (i.e., exponential, logarithmic, etc.), though the general trend is indicated by the linear fit. Perhaps a more appropriate scale for the x-axis is the Bark scale, which measures the perceived frequency. For the range 1400 to 1900 Hz, however, the Bark to Hertz transformation is very close to linear in this range and therefore does not affect the linear fit seen in Figure 15.

3.4 Summary and discussion

Listener stop identification rates in a perceptual experiment ranged from 79.4% to 96.2% for all CV contexts. As in previous studies, the velar stops caused greater listener errors than other stop places. Stops preceding the [i] context, however, did not have significantly lower identification rates than stops in other contexts. Listener error rates by CV pair correlated to DT classification errors by CV pair, indicating that listener performance in stop place identification depends at least in part on the amount of discriminatory information in the acoustic signal.

Individual features used by listeners to identify stop place were identified in almost all CV contexts using stepwise linear regression (Table 6). For each of the four CV contexts that showed the greatest difference in listener error rates between canonical and non-canonical CVs ([pu], [ta], [ki], and [ku]), one or more acoustic features predicted by DT classifications to be relevant to stop place identification was found to significantly affect listener responses. In all cases, the range of values for each significant feature that corresponds to the highest identification rates in the perception study is the same range of values considered characteristic for the given CV token based on feature values of all CV tokens collected, suggesting that listeners use knowledge about the acoustics characteristics of naturally varying tokens in the perception of stop place.

The acoustic features found to significantly affect listener response rate varied by CV context. In most cases, the features found to significantly affect listener response rates were among those features predicted by DT classifications. The relative ranking of the single-feature DT classification accuracies, however, was not reflected in listener response rates.

Stimulus	% Correct overall	Significant feature(s)
[tu]	96.2	VOT
[pa]	94.7	TRI_ERR
[pi]	93.2	REL_AMP
[ti]	90.8	---
[ta]	87.4	VOT, LIN_SLPE
[pu]	87.3	F2_10MS, LIN_SLPE
[ku]	86.1	NODE, F2_00MS, VOT (?)
[ka]	84.2	VOT
[ki]	79.4	VOT, F2_10MS (?)

Table 6. Percentage of correct listener responses by CV and the acoustic features responsible for stop place identification. Results are sorted by identification rates. '?' indicates a near-significant correlation.

VOT was found to significantly affect listener response rates for five CV contexts. This result suggests that VOT has been largely overlooked as a cue for stop place, though it is considered a primary cue to manner (voiced, voiceless unaspirated, aspirated). It is also possible that the relatively direct method of VOT extraction from the signal gives it an advantage in linear regression models.

The linear regression correlations in all cases were relatively weak, with R^2 values ranging from 0.2 to 0.6. Indeed, one drawback of using natural data and investigating a large set of cues at one time is that feature effects cannot be entirely factored out or controlled for. Although linear regression models are able to capture significant effects of particular acoustic features on listener accuracy, they do not constitute a complete explanation for listener confusions. On the other hand, listeners are known to use as many cues as are available to them, including non-phonetic cues, such as phonotactic frequency. The correlations between listener accuracy and predicted feature values show that at least some portion of listener confusions are due to inherent properties of the acoustic signal and listeners' knowledge of how phonemes in different contexts are distinguished along these properties. In most cases, listeners rely on those acoustic cues that are the most informative in a given context.

Not all acoustic features predicted by DT classifications to be inherently informative were used by listeners in stop place identification. The feature REL_AMP, for example, was one of the best features for distinguishing [p] from other stops, especially [t] in the context of [u], based on properties of the acoustic signal of all CVs studied. The percentage of correctly identified [pu] stimuli in the perception experiment, however, showed no relation to their values of REL_AMP (Figure 16).

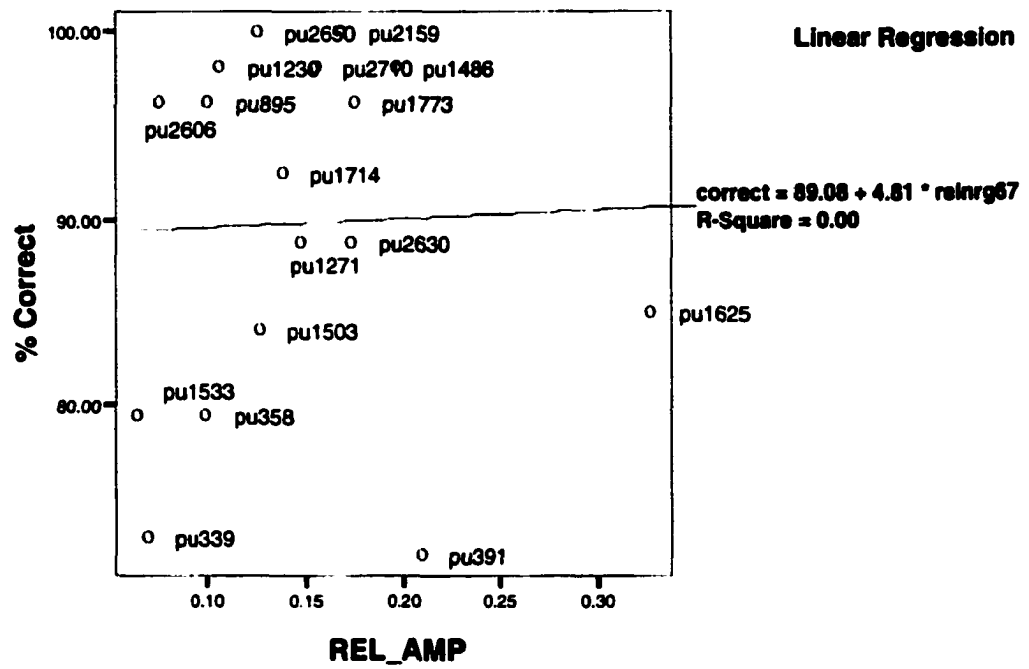


Figure 16. Linear regression of percentage of correct responses to [pu] stimuli by the REL_AMP values of the stimuli.

4 Conclusion

Results from the perception study presented in this chapter support the claim that listeners are more likely to make stop place confusions when a given context offers fewer cues: The percentage of listener error by CV pair significantly correlated with CV-specific accuracies of multiple-feature DT classifications by CV pair. A multiple stepwise linear regression model found several acoustic features predicted by the DT classifications in Chapter 3 to be responsible for listener response rates. Listeners were

more likely to correctly identify a particular CV stimulus the closer that stimulus lay to the mean along the relevant feature (or features).

An alternate approach to comparing human and DT performance would be to use human categorizations of acoustic segments as the basis for classification, in lieu of the CV intended by the speaker. If DTs were trained on the associations between acoustic properties and human categorizations of speech sounds, the resulting trees would indicate a more accurate relative ranking of acoustic features with respect to human judgments of stop place. For example, if humans are more likely to use F2 onset than spectral shape as a cue to stop place, to the extent that they miscategorize CVs that have non-canonical F2 onsets, trees trained on human categorizations would use F2 onset as a primary feature for splitting data. We leave this approach to further studies.

In the following chapter, I concentrate on the contexts and features responsible for the directionality of errors. Using the same linear regression techniques, I will determine what individual features are responsible for specific stop place confusions ([ki] → [ti], for example). The investigation into the possible source of unidirectional confusions begins with an examination of several hypothesized causes, including the frequency of CV tokens in the lexicon, the distributional shape of CV tokens along relevant feature values, and physical properties of the specific acoustic features involved.

Chapter 5

Directionality of perceptual errors

1 Introduction

In Chapter 4, the rates of correct responses in a perception study were examined for any correlation with featural and contextual salience of stop place. In this chapter we turn our focus to rates of *errors* in the perception study, which share patterns with many previous confusion studies (Chapter 1, Section 5). In particular, we will investigate possible explanations for asymmetric confusions (in which stop A is often confused for stop B but stop B is rarely confused for stop A, noted A → B), several of which parallel historical sound changes. Previous explanations relying on notions of *similarity* correctly point to pairs of phonemes that are most likely to cause errors in listener identification tasks, but they fail to provide an adequate explanation for why confusions would be asymmetric. Similarity between two stops alone would predict that each stop is equally likely to be confused for the other.

In this chapter, I begin by reviewing a number of explanations put forth for the asymmetry of stop place confusions. These explanations include the proposals that consonants may have *affinities* for certain vowels (Repp & Lin, 1989), that error patterns may be linked to the frequency of a segment in the lexicon (Vitivitch 1999), that certain segments may be *marked* due to production constraints (Lindblom 1986, 1990), and that degradation of non-robust features, which is inherently asymmetric, may lead to

perceptual asymmetries (Plauché et al. 1997; Chang et al. 2001). Finally, I present a fifth possibility: Listeners may disfavor stop places associated with greater token-to-token variation along a primary cue than those with less variation.

2 Stop consonant and vowel affinities

Repp & Lin (1989) found that in a consonant identification task, alveolar consonant identification of CVs was improved when the vowel was fixed, especially in front vowel contexts, whereas the labial and velar consonant identification improved in front vowel contexts but was impaired in back vowel contexts. They proposed that some vowels may have *affinities* with certain consonant places due to their articulatory and spectral properties. This would account for response biases for bilabial stops in the context of rounded back vowels and for alveolar stops in the context of high front vowels. The response biases are expected only when the vowel context is unknown, however. According to this account, errors in stop place will be more evenly distributed across the three stop places if the vowel is known. For example, they found that in English, knowledge of the vowels increased identifiability for [k] and [p] when followed by [u], and especially [k] when followed by [a].

In the current study, both [p] and [k] stimuli are commonly confused for alveolar place of articulation in the context of [i]. The trend mentioned by Repp & Lin for [t] and [k] to be confused for [p] in the context of [u] was not found, however. In addition, velar stops caused the highest rates of confusions, but they did so across all vocalic contexts.

We can test whether response biases are expected only when the vowel context is unknown by comparing the error rates in the two versions of the perceptual experiment.

In version A, the vowel of the CV stimuli was varied randomly throughout all blocks. In version B, the stimuli were presented in vowel-specific blocks. Version A is predicted to show relatively high errors for bilabials and velars in the [u] and [a] contexts as well as for alveolars in the [i] context. Version B, on the other hand, should show errors that are more evenly distributed across the three stop place categories.

In a one-way ANOVA analysis, the response rates across all CV contexts did not vary significantly by the version of the test ($F(1, 15728) = 0.016$). Individual CV contexts showed very little variation in incorrect responses between version A and version B of the experiment. *T/D* responses in the context of [i] were not found to be significantly more likely in version B than in version A (Figure 1, 3), nor were *P/B* responses in the context of [u] more likely in version B than in version A (Figure 2, 3). Instead, [pu] stimuli were confused for alveolar place of articulation more often when the vowel was known than when it varied randomly (Figure 1). This was also true to some degree with [pu] confusions for velar place of articulation.

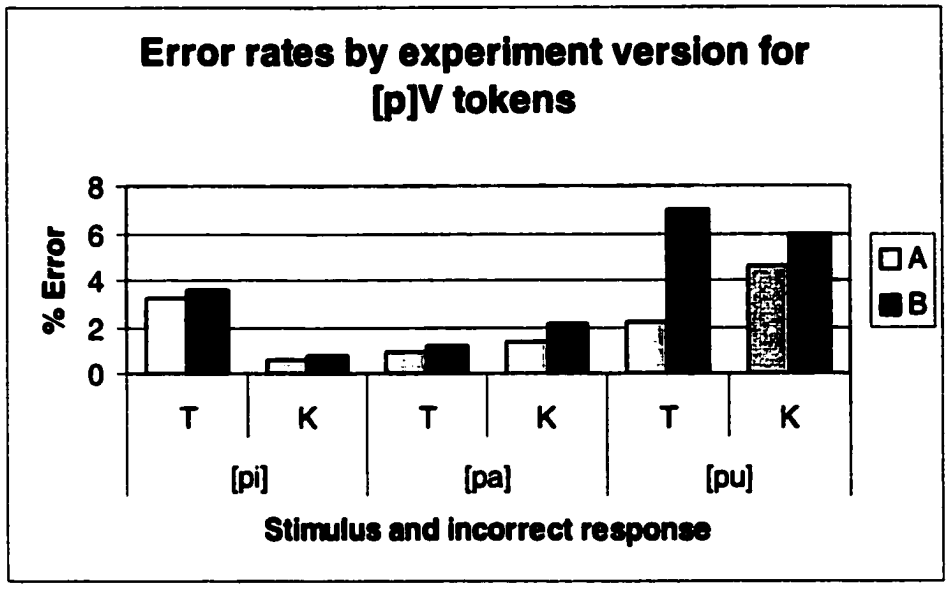


Figure 1. Stop place confusions for [p]V tokens. Version A consisted of randomly varying vowels. In version B, the vowel of the CV stimuli was known.

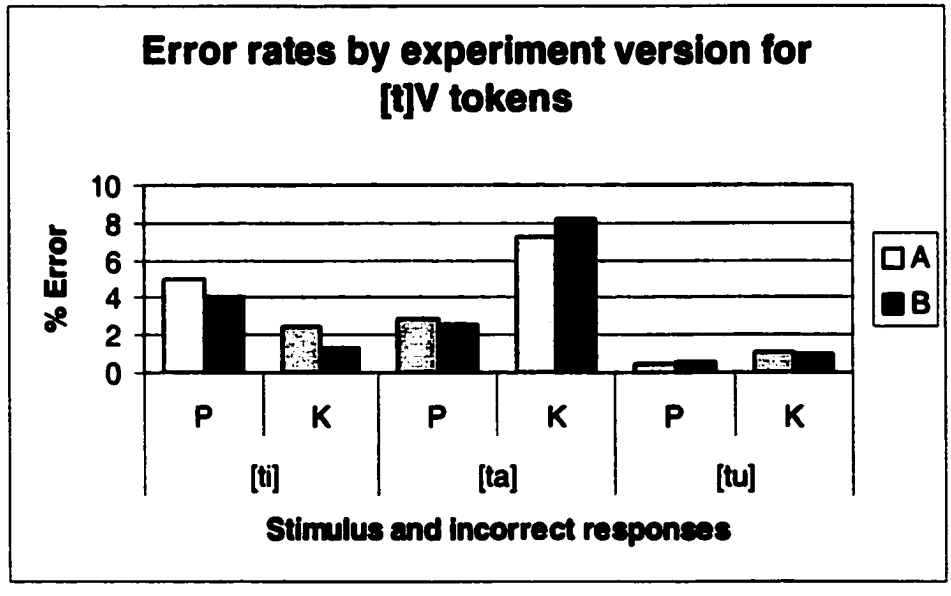


Figure 2. Stop place confusions for [t]V tokens. Version A consisted of randomly varying vowels. In version B, the vowel of the CV stimuli was blocked.

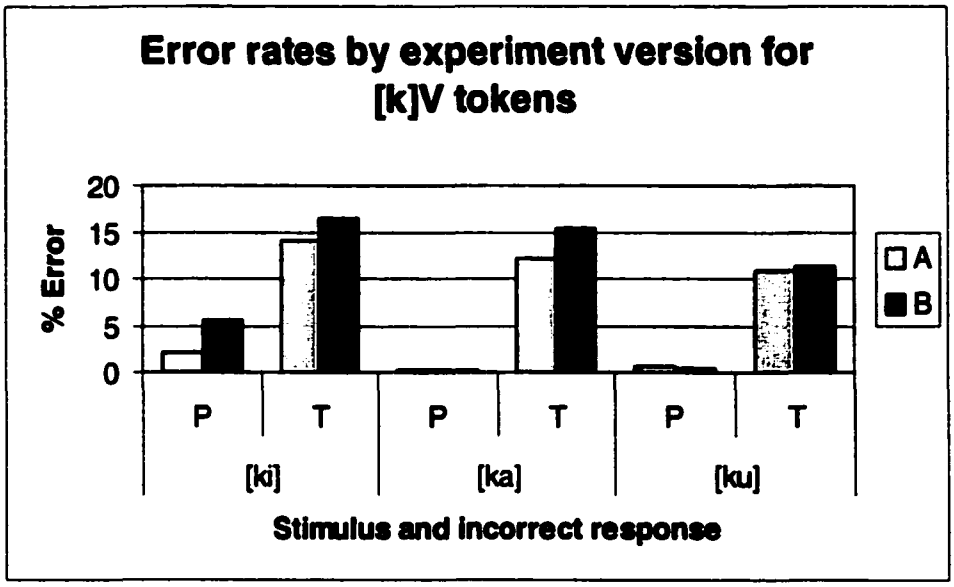


Figure 3. Stop place confusions for [k]V tokens. Version A consisted of randomly varying vowels. In version B, the vowel of the CV stimuli was blocked.

Overall, the results in this experiment did not support the claim that affinities between vowels and consonant account for response biases toward alveolars preceding high front contexts and bilabials preceding back rounded vowels. In addition, this claim offered no explanation for the most prominent asymmetric confusions, [k] → [t] in all contexts.

3 Frequency of segments in the lexicon

Another explanation often offered for asymmetries in stop consonant confusions lies in the frequency of the phoneme in the language (Lyublinskaya 1966). Vitevitch et al. (1999) found that in word recognition tasks, non-words with frequent phonotactic patterns in the language are likely have more rapid and more accurate identification than

non-words with infrequent phonotactic patterns. This result suggests that certain CVs may be protected from confusions by virtue of their frequency in the lexicon.

The frequency of all CV phonotactic patterns in English (where C = {p, t, k, b, d, g} and V= {i, a, u}, stressed) was examined in two representative English corpora: the Brown corpus (Kucera & Francis 1967) and the Switchboard Corpus (Godfrey et al. 1992) (Table 1, Figures 4 and 5). The Brown corpus is a corpus of written American English collected from edited prose from a variety of materials published in 1961. The Switchboard Corpus is a corpus collected from transcriptions of telephone conversations between strangers on pre-arranged topics that is considered to represent near-natural speech. For each CV segment, the number of word *types* (word entries in the corpus that include the pattern in question) and the number of word *tokens* (actual instances of words containing the target pattern) are reported out of all types and tokens with any singleton onsets in pre-stress position. Flaps (e.g., the [ɾ] in *city* [sɪ́ɾi]) were not included in the CV patterns examined.

		Switchboard Corpus		Brown Corpus	
CV		Types	Tokens	Types	Tokens
[p]/[b]	-[i]	**245** (11.4 %)	**29563** (14.8 %)	**305** (10.4 %)	**19448** (17.2 %)
	-[a]	**363** (16.9 %)	**11845** (5.9 %)	**528** (18 %)	**8327** (7.4 %)
	-[u]	**62** (2.9 %)	**375** (0.2 %)	**72** (2.5 %)	**358** (0.3 %)
[t]/[d]	-[i]	**297** (13.8 %)	**10792** (5.4 %)	**434** (14.8 %)	**26562** (23.5 %)
	-[a]	**259** (12 %)	**8394** (4.2 %)	**335** (11.4 %)	**4140** (3.7 %)
	-[u]	**201** (9.4 %)	**109646** (54.8 %)	**278** (9.5 %)	**35817** (31.7 %)
[k]/[g]	-[i]	**77** (3.6 %)	**2596** (1.3 %)	**84** (2.8 %)	**7176** (6.4 %)
	-[a]	**566** (26.3 %)	**22931** (11.5 %)	**817** (27.8 %)	**9972** (8.8 %)
	-[u]	**79** (3.7 %)	**4091** (2 %)	**85** (3 %)	**1090** (1 %)
Total CVs in Corpus		**2,149**	**200,233**	**2,938**	**112,890**
Total words in Corpus		**26,574**	**3,438,550**	**43,020**	**1,014,312**

Table 1. Frequency of CV tokens in the Brown and Switchboard corpora.

In the Switchboard corpus, out of 3,438,550 uttered words from 26,574 possible word types, only 2,149 words (200,233 tokens) contained the phonotactic pattern CV, where C = {p, t, k, b, d, g} and V = {i, a, u}, stressed. The percentages by CV segments show that [ka] and [ga] are found in the most word types (Figure 4), but that the segments [tu] and [du] were uttered the most frequently (Figure 5). The large number of [tu] and [du] tokens was due to the frequency of the word *do* in both written and spoken English.

Although there are twice as many possible word types with a CV sequence in the Brown corpus as in the Switchboard (43,020 as compared to 26,574), the distribution of those word types by CV segment is remarkably similar across the two corpora (Figure 4). In the Brown corpus as well, word types containing the sequences [ka] or [ga] were more common than words containing other CV sequences. The frequency of [ka] and [ga] sequences are primarily due to the common occurrence of the words *god, got, call, car, common, costs,* and *cause* in spoken and written English.

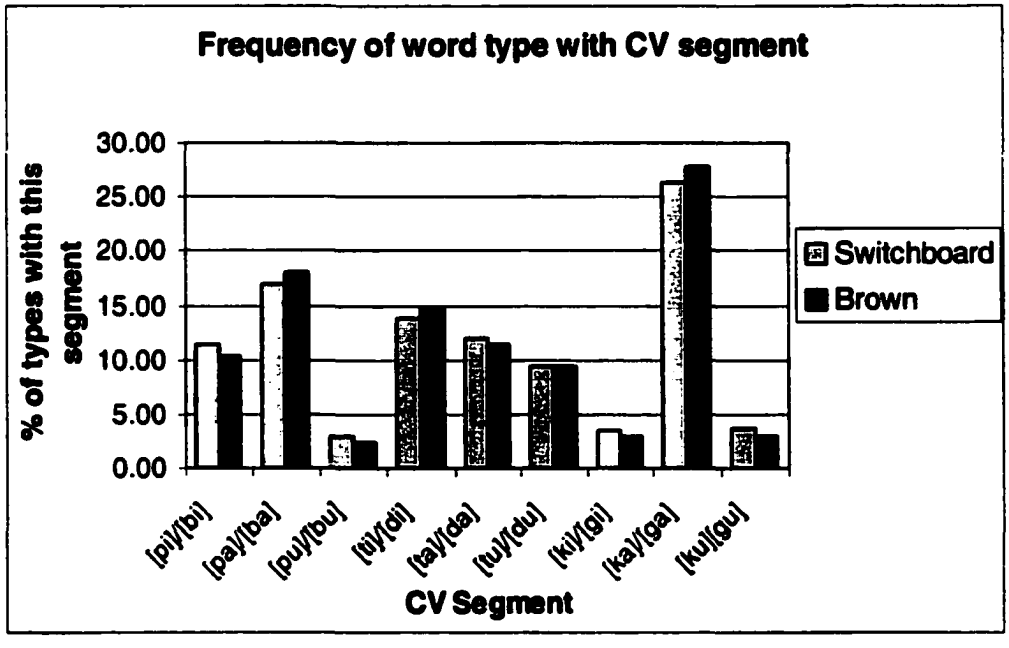


Figure 4. Word types in Switchboard and Brown corpora containing CV segments. The Y-axis represents the percentage of CV tokens where C={p, t, k, b, d, g} and V={i, a ,u}, stressed.

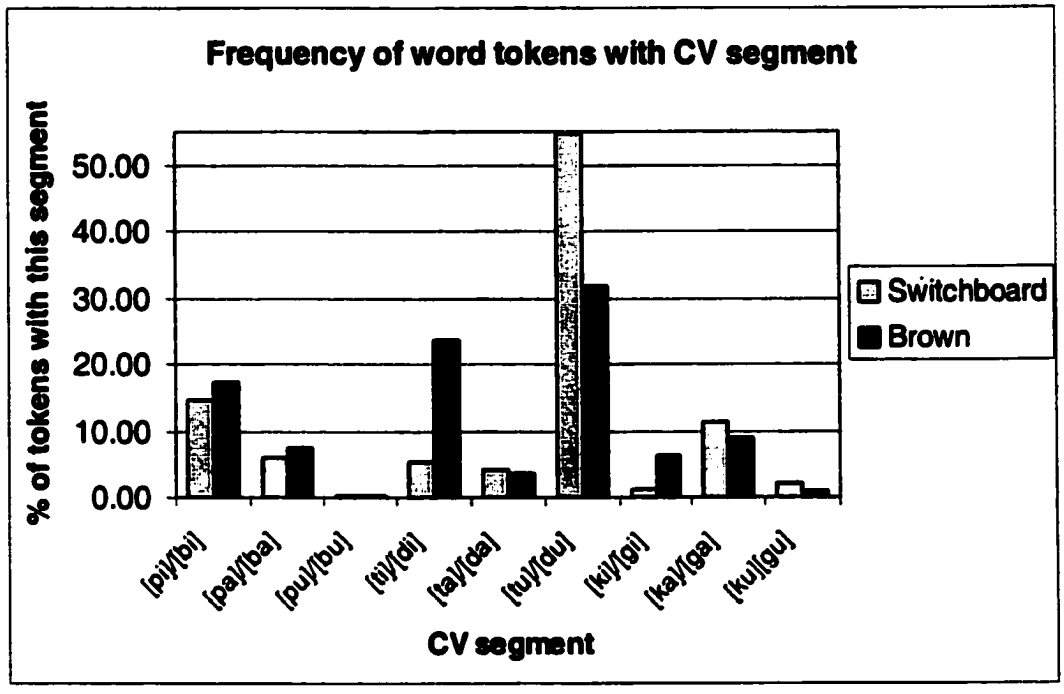


Figure 5. Utterances in Switchboard and Brown corpora containing CV segments. The Y-axis represents the percentage of CV tokens where C={p, t, k, b, d, g} and V={i, a ,u}, stressed.

Alveolar stops preceding stressed vowels occur in more spoken and written English words than the corresponding bilabial and velar stops. The account offered by Vitevitch et al. (1999) – that these segments are more accurately identified by listeners because of their higher frequency in the language – appears to hold, since confusions for alveolar place are more common than for any other stop place. A significant correlation exists between the frequency of word tokens in the Brown corpus containing a given CV segment and percentage of times the CV is given as an incorrect response in the perception experiment ($R^2 = 0.46$, $p < 0.005$) (Figure 6). A similar correlation was found for the frequency of a given CV segment in the Switchboard corpus ($R^2 = 0.34$, $p <$

0.005). CVs that are more frequent in a listener's lexicon are more likely to be given as incorrect responses in the stop place identification task.

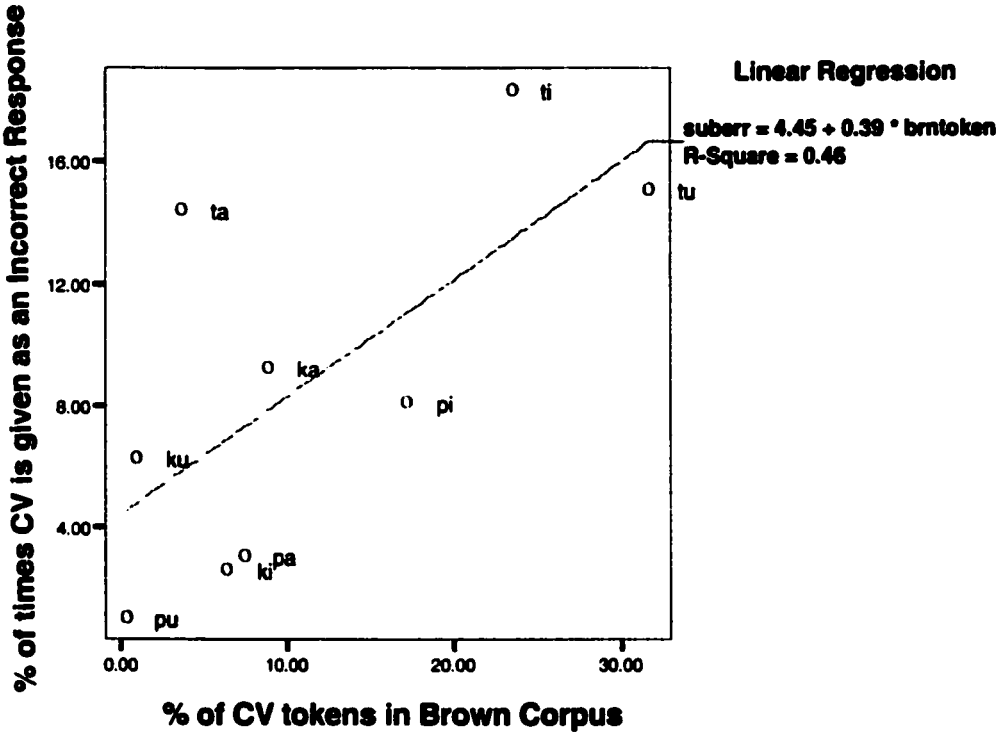


Figure 6. Percentage of times a CV is given as an incorrect response by the frequency of the CV in the Brown corpus.

Asymmetries reported in previous confusion studies, such as [ki] → [ti] and [ku] → [tu], correspond to asymmetries in CV frequencies; [ti] is two to three times more frequent than [ki] in the CV context. Additionally, a frequency account would predict that [ki] to [pi] confusions would be more common than [pi] to [ki] confusions, since the segment [pi] appears in both written and spoken speech at far greater frequencies than [ki]. In the perceptual experiment summarized previously, the latter finding is true: [pi]

is confused for [ki] a mere 0.7%, while [ki] is confused for [pi] 3.6% of the time. On the other hand, the perception study showed that [ka] was confused for [ta] at high rates but the reverse was not true. The frequencies of these segments would predict the opposite direction, or none at all, since [ka] is found in more written and spoken word types than [ta] and is produced more frequently in both corpora.

In contrast to explanations that focus on the cause of listener errors, Vitevitch et al. (1999) claim that the frequency of a phonotactic pattern *facilitates* the identification of a that pattern. Therefore, some ambiguity at the level of the acoustic signal is still required to trigger confusions. Given an ambiguous signal, however, CVs that are frequent in the listener's lexicon are less likely to be mistaken for another stop place. The following sections examine ambiguity in the signal and its role in the directionality of stop place confusions.

4 Markedness vs. acoustic properties

Many accounts for asymmetries in stop confusions view certain stops as *marked* and others as *unmarked.* A marked phoneme, such as [p] or [k], is more likely to be confused for an unmarked or default phoneme, such as [t]. Markedness is defined as any non-acoustic considerations that listeners employ in the task of identification or differentiation. These considerations may include an SPE-based (Chomsky & Halle 1968) universal disfavoring of certain sounds founded either on auditory distinctiveness or articulatory ease (Lindblom 1986) or a ranking of sounds based on phonological patterning, for example. Under these definitions of markedness, [p]'s and [k]'s are argued

to be universally more marked than alveolar stops based on cross-linguistic frequency or articulatory or perceptual factors.

Lindblom (1986) suggests that the interaction between *auditory distinctiveness* and the *extremeness of articulation* generates many patterns described as markedness relations between different segments. A language must find a balance between a phoneme's perceptual salience and the extremeness of the movement of the articulators involved in its production. Generally, Lindblom claims, extremeness of movement outweighs the perceptual salience of gesture. Thus, though [gi] may be more salient than [gu], [gu] is favored for its lower physiological cost. Speakers presumably learn these patterns of production and apply them to perception.

Chang et al. (2001) showed that at least in the case of [ki] → [ti] confusions, though markedness may cause a response bias for alveolar place of articulation, the acoustic and auditory characteristics of this pair of CVs are the primary trigger for the asymmetry. The first clue to the possible cause for [ki] → [ti] confusions came from a visual perception study in which the subjects were asked to identify Roman capitalized letters (Gilmore et al. 1979). The following asymmetric confusions were found:

E → F

Q → O

R → P

In visual perception tasks, *E* was often confused for *F* but *F* was rarely confused for *E*. Likewise, *Q* was often confused for *O*, but *O* was rarely confused for *Q*, and so on. In each case, the pairs of letters are structurally similar with the exception of an *extra* feature that the letter on the left possesses and the letter on the right lacks. Subjects are

thought to be more likely to accidentally miss the extra feature of the letter on the left, resulting in the confusions shown above, than they are to accidentally *see* the extra feature in the letter on the right.

The same concept may be applied to the auditory domain. As discussed in Chapter 3, the acoustic cues to stop place show differences in their inherent salience depending on the CV context. All acoustic cues are sooner or later subject to degradation during their transmission from speaker to listener. Those cues that are more salient or more robust are less likely to be degraded to the point of losing their distinctiveness than less salient cues. The degradation of a non-robust cue may lead to a perceptual confusion, especially when no other cues are available. According to the second law of thermodynamics, entropy (the degree of disorder in a system) never decreases within a domain without the input of more energy, which is not applicable here. Thus, degradation of the signal during transmission is not expected to lead to an increase in the distinctiveness of acoustic cues.

In the case of [ki] → [ti] confusions, Plauché et al. (1997) found that due to the raised F2 of the following [i], the formant transitions of [ki] and [ti] are structurally similar. The spectra of the stop bursts were also similar, with the exception that [ki] has an *extra* non-robust mid-frequency spectral peak. They hypothesized that if the mid-frequency peak is degraded enough, it could lose its contrastive function and listeners would be more likely to confuse [ki] for [ti]. At the same time, listeners would rarely confuse [ti] for [ki], despite the structural similarities, since they would be unlikely to spuriously insert the mid-frequency peak. And indeed, when the mid-frequency spectral

peak of [ki] tokens was degraded using a band-reject filter, [ki] → [ti] confusions greatly increased (Plauché et al. 1997).

The acoustic/auditory explanation was shown to override any possible markedness effects by Chang et al. (2001) in two experimental studies. In the first, despite any possible response bias favoring [ti] over [ki] due to markedness (see Section 3), confusions can be induced for some listeners in the opposite direction ([ti] → [ki]) by bandpass-filtering white noise between the center frequencies (2880 to 3880 Hz) and mixing the filtering noise with the burst of the alveolar stop in [ti].

In a second study, the [ki] → [ti] asymmetry is shown to be specific to the high, front vocalic environment. The mid-frequency region of velar bursts was bandreject-filtered in the following contexts: -[i], -[au], -[er], -[u], -[ei], -[aɪ], -[æ], -[a], -[ou]. Filtered velar stops preceding high front vowels [i] and [ei], as well as the front vowel [æ], caused the most confusions. Filtered velar stops preceding other vowels did not cause confusions for alveolar place, even though alveolars are presumably unmarked in all vocalic contexts. This result supports the acoustic explanation for the [ki] → [ti] asymmetry and challenges the markedness account, according to which velars are universally more marked than alveolars regardless of the quality of the following vowel. Interestingly, in the current study, velars in all three vocalic environments were frequently confused for alveolar place.

Extra features of non-robust acoustic properties that are responsible for individual asymmetric confusions must be investigated on a case-by-case basis as described above for [ki] → [ti]. In the following section, acoustic features that are potentially relevant for asymmetric confusions are identified for further study. In addition, a final hypothesis –

that asymmetric confusions stem from asymmetries in distribution of CV tokens along relevant acoustic features – is investigated.

5 Asymmetries at the feature level

In Chapter 4, DTs, which rely on distributions of stop categories along acoustic features were found to correlate significantly with human perception of stop categories. Features along which a given pair of stops shared few values resulted in very accurate DT classifications of stop place. Features along which a given pair of stops shared many values were not useful in stop place discrimination and resulted in low DT accuracies. What about cases where differences in distribution exist *within* a pair of stops?

In this section we investigate the possibility that differences in distribution within a pair of stops along feature values may favor one stop category over the other in listener's categorization of stop place. The hypothesis assumes that listeners make categorical decisions based on the most likely candidate for stop place, much like a Bayesian classifier.

Imagine that you hear the sound of your garbage can being overturned one evening and you know from previous experience that your neighbor's dog or cat is likely to be the culprit. By the time you reach the curb, however, you can detect only a dark form at the end of the block. Under better circumstances, you might use any number of visual cues to identify the type of animal, including: size, coloring, movement, shape and relative size of ears, snout, tail, etc. On this particular evening, the animal is still and a mere silhouette; you must rely on only its size for identification. We know that dogs vary greatly along this feature, ranging from the size of a breadbox to roughly the size of a

small horse. Cats, however, vary much less in size by individual, ranging from the size of a muffin to the size of a breadbox. The distributions of cat and dog populations by their approximate size are shown in Figure 7.

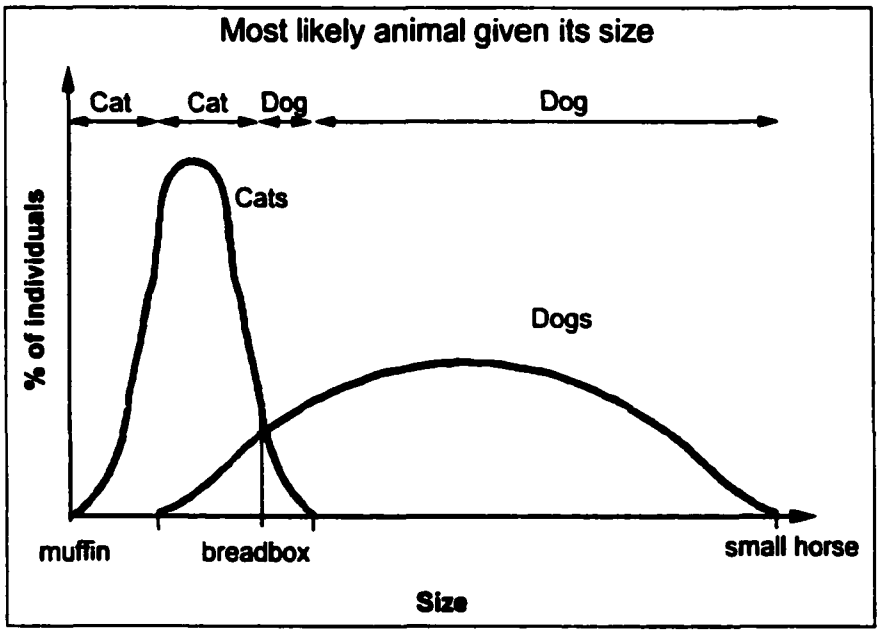


Figure 7. Distributions of cats and dogs by size. Sizes are approximate.

Assuming that the animal at the garbage cans is equally likely to be a cat or a dog (i.e., equal priors). If the animal is smaller than a breadbox (but larger than a muffin), it's a cat. If the animal is larger than a breadbox, it's a dog. If the animal is about the size of a breadbox, it could be either. If humans are like Bayesian classifiers, making categorical decisions by estimating the greatest probability for each category, an animal about the size of a breadbox is more likely to be a cat, based on what we know about the distribution of cat and dog populations by their size. Only a small percentage of all dogs are the size of a breadbox, but the majority of all cats are close to that size. Thus, if you

must rely on the size of the animal as the primary visual cue to its identity, in ambiguous cases (i.e., when the animal is about the size of a breadbox), you are more likely to confuse a dog for a cat than you are to confuse a cat for a dog.

In this section, we test the hypothesis that identification based on Bayesian probabilities such as the example above lead to asymmetries in listener's categorization of stop place. A non-parametric method is used to estimate the informativeness of individual features, based on the relative distribution of stop categories along feature values. Differences in relative informativeness between a pair of stops are then tested for their role in listener errors.

5.1 Non-parametric estimation of percentage overlap

Histograms of pairs of stops along individual features reveal differences in the overall distribution of stop place along given feature values. In particular, they indicate value ranges for which the majority of tokens belong to a single stop place (*majority region*) and value ranges for which the tokens belong to more than one stop place (*overlap region*). For example, a histogram of [pa] and [ta] tokens by their value of F2_10MS reveals a majority [p] region, a majority [t] region, and an overlap region (Figure 8). The differences in the overall shape of the distribution of [pa] tokens and [ta] tokens over F2_10MS are also apparent.

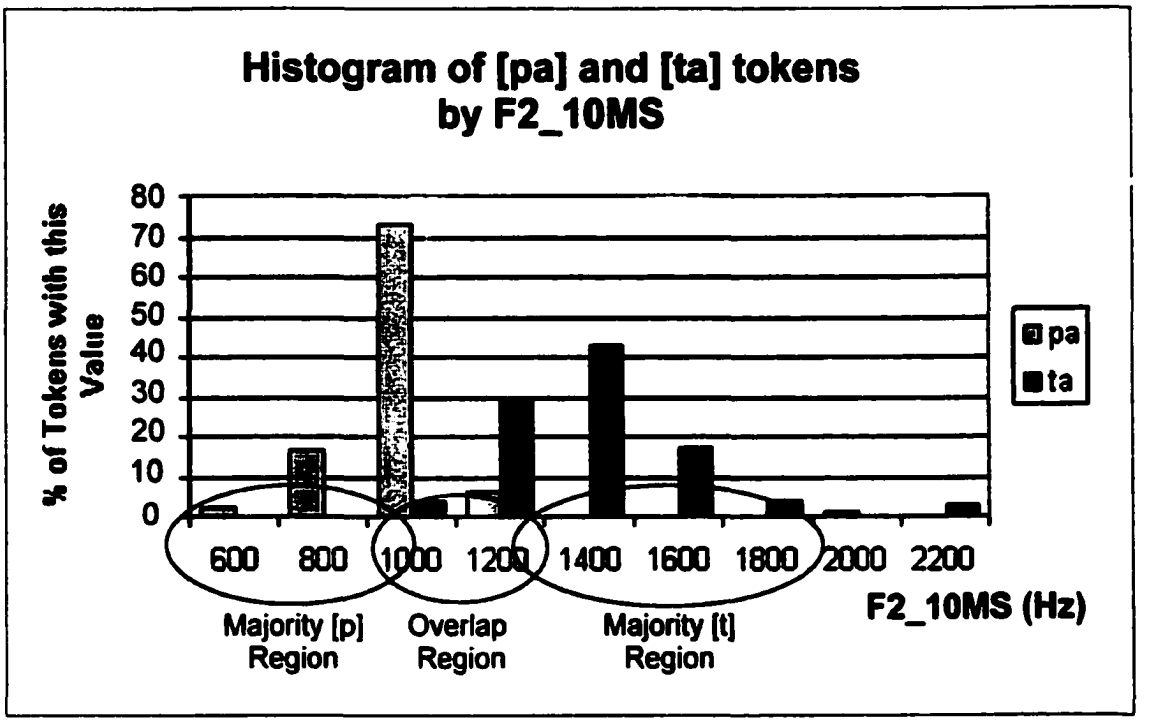


Figure 8. Histogram of [pa] and [ta] tokens by F2_10MS. The feature F2_10MS is binned with a bin size of 200 Hz.

The graph shows that most of the [p] tokens have F2_10MS values between 1000 and 1199 Hz. The [t] tokens in the [a] environment, however, mostly range from 1200 to 1799 Hz. The small degree of overlap between the values for [pa] and [ta] suggests that the value of F2_10MS contains information about the place of articulation of the stop consonant that is available to listeners for discriminating between [p] and [t] in the context of [a]. Recall that the DT classification using this feature was near perfect (94.44%). When stimuli in the overlap range are presented to listeners, neither [pa] nor [ta] responses are expected to be favored.

Now we turn to a histogram of the same [p] and [t] tokens in the same environment, the vowel [a], but this time illustrating the distribution of these tokens along

the values of LIN_ERR, the error of the linear fit to the spectrum at the stop burst (Figure 9). Here the majority of both [p] and [t] tokens lie between the values of 0 to 1200 Hz. There is no value range for which [p] tokens represent the majority of tokens, though values above 3000 Hz are attributed only to [ta] tokens.

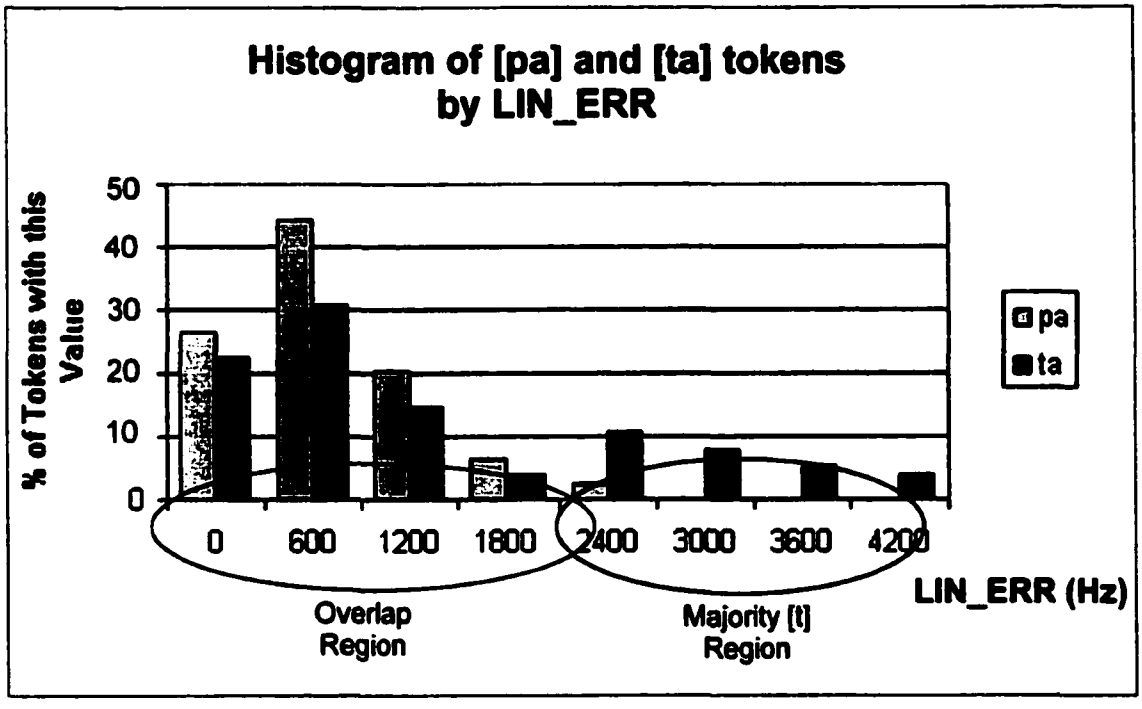


Figure 9. Histogram of [pa] and [ta] tokens by LIN_ERR. The feature LIN_ERR is binned with a bin size of 600 Hz.

The high degree of overlap between the [pa] and [ta] tokens suggests that this acoustic feature contains very little information for discriminating between [p] and [t] in the [a] context on its own, though this same feature may be useful in combination with another feature, in another vowel environment, or for another pair of stops. Note also that the *majority [t] region* from 3000-4799 Hz does not contain the majority of [ta]

tokens, since most [ta] tokens fall in the same range as the [pa] tokens, from 0 to 2999 Hz. Recall that the accuracy of DT classification of [pa] and [ta] along this feature was less than 75%. Based on the shapes of the [pa] and [ta] token distributions alone, for stimuli that lie in the overlap region, [pa] responses are favored, since [pa] tokens represent the majority of all tokens in the overlap region. The non-parametric method discussed below will capture this asymmetry at the feature level.

The amount of overlap in the two histograms (Figures 8 and 9) can be estimated using a simple rank-based non-parametric method, which makes no assumptions about the shape of the CV distributions. As in the case of decision trees, this method is sensitive to outliers, but allows more control over their effect, since the overall distribution information is available in the histogram. For the purposes of this study, a strict criterion for outliers was adopted: Boundary values separated from the rest of the data by a single 0-valued bin or several contiguous 0-valued bins are considered outliers if their removal results in a histogram that is monotonic on either side of the peak value. The total number of tokens is adjusted after any outliers are removed. This method works well only if the width of the bins is relatively large with respect to the range of values possible, as it is in this study.

Using the histogram of [pa] and [ta] tokens along values of F2_10MS as an example (Figure 8), the algorithm for calculating the percentage of overlap works as follows:

1. *The lowest [t] value is identified.*

In this case: ta $_{F2\_10MS}$ (1)= 1226 Hz.

2. *The lowest [t] value is compared to the list of [p] values. The number of [p] values that come before this value represents the [p] quantile at which the [t] value starts.*

In this case: There are 78 [p] tokens with values lower than 1226 Hz (Table 2).

Quantile	Sorted list of [pa] F2_10MS values	
:	:	
:	:	
75	1180.99	
76	1187.6	
77	1189.81	
78	1199.93	
79	1240.58	← 1226.08 (lowest [ta] value)
80	1240.88	

Table 2. Sorted list of [pa] F2_10MS values with respect to the first [ta] value.

3. *The quantile at which the [t] value starts is subtracted from the total number of [p]'s, then divided by the total number of [p]'s. This yields the percentage of the total [p] values that lie in the overlap range.*

In this case: % of [p]'s in the overlap range = (79-78)/79 = 1/79= 1.2%

4. *The highest [p] value is identified.*

In this case: pa $_{F2\_10MS}$(79) = 1241 Hz

5. *The highest [p] value is compared to the list of [t] values. The number of [t] values that fall after this value represents the [t] quantile (from the end) at which the [p] values stop.*

In this case: There are 73 [t] tokens with values greater than 1241 Hz (Table 3).

Quantile	Sorted list of [ta] F2_10MS values	
1	1226.08	
2	1239.1	
3	1271.53	← 1240.88 (highest [pa] value)
4	1281.51	
5	1293.24	
6	1309.77	
⋮	⋮	

Table 3. Sorted list of [ta] F2_10MS values with respect to the last [pa] value.

6. *The number of [t]'s greater than the highest [p] value is subtracted from the total number of [t]'s, then divided by the total number of [t]'s. This yields the percentage of the total [t] values that lie in the overlap range.*

In this case: % of [t]'s in the overlap range = 75-73/75 = 2/75 = 2.6%

The same algorithm is applied to the histogram of [pa] and [ta] distributions along LIN_ERR (Figure 9). In this case, only one [p] token has a value lower than the lowest [t] value (109 Hz). By the algorithm above, the percentage of [p]'s in the overlap range is 98.7%. There are seven [t] tokens with values higher than the highest [p] value (2531 Hz) along the feature LIN_ERR. This yields a percentage of [t]'s in the overlap range of 77.3%.

The differences in the amount of overlap of [pa] and [ta] distributions along features F2_10MS and LIN_ERR seen in Figures 8 and 9 are captured by the non-parametric method for calculating the percentage of tokens in the overlap range. The results from applying this algorithm to the two cases are summarized in Table 4.

% Overlap by feature	F2_10MS	LIN_ERR
% of [pa] in overlap range	1.2	98.7
% of [ta] in overlap range	2.6	77.3

Table 4. Summary of % of overlap for [pa] and [ta] for features F2_10MS and LIN_ERR.

The percentage of tokens in the overlap range for these two features captures interesting asymmetries in CV distributions by feature. For the LIN_ERR, for example, almost all of the [pa] tokens share values with certain [ta] tokens. Only 77.3% of [ta] tokens, however, lie in an overlap region; more than one-third of all [ta] tokens have LIN_ERR values that are greater than those of any [pa] tokens. Stimuli that lie in the overlap range may be more likely to be heard as [pa], since the majority of [pa] tokens lie in the overlap range. Stops that show less variation (exhibit narrow-band histograms) will have higher percentages of overlap than stops with the same mean value but with more variation (exhibit broad-band histograms). In this section we will investigate whether differences in distribution may lead to asymmetries in listener perception.

If listeners are aware of the relative distributions of stop place along relevant features, in cases of ambiguity (tokens in the overlap range), they are hypothesized to confuse a stop with a lower percentage of tokens in the overlap region for a stop with a higher percentage of tokens in the overlap region. In this section we will examine the significant asymmetric confusions to determine whether the direction of stop place confusions correlates with the relative difference in percentage overlap between the two stops along a relevant feature.

5.2 Percentage overlap in asymmetric confusions

The confusion matrix from Chapter 4 is replicated below with the significant stop place confusions in bold (Table 5). The following stop place confusions are investigated for any responsible acoustic features that may have triggered them: [ki] → [ti], [ku] → [tu], [pu] → [tu], [pu]→[ku], and for comparison, the symmetric confusion [ka] ←→ [ta].

Stimuli / Response	[p]			[t]			[k]		
	[i]	[a]	[u]	[i]	[a]	[u]	[i]	[a]	[u]
P/B	93.2	94.7	87.3	4.5	2.8	0.5	3.6	0.3	0.6
T/D	3.4	1.0	**4.3**	90.8	87.4	96.2	**15.0**	**13.5**	**10.8**
K/G	0.7	1.7	**5.2**	2.0	**7.6**	1.1	79.4	84.2	86.1
NA	2.7	2.5	3.2	2.7	2.2	2.3	1.9	2.0	2.5

Table 5. Confusion matrix (from Chapter 4, Table 3). *NA* indicates a *no answer*, in which the subject took longer than 3 seconds to make a decision. Confusions of interest are shown in bold.

In this study, we are interested in determining what values of a given feature are needed to push the listeners' percept to an incorrect place category. For this purpose, the raw feature values of CV stimuli are not useful. Instead, the rank normalized value of a token with respect to the target category is used in place of raw feature values in the following analysis. That is, each token of stop A has a rank with respect to the mean of stop B for a given feature. If stop A has a smaller feature value than the mean of stop B, the rank has a negative value. If stop A has a greater feature value than the mean of stop B, the rank has a positive value. The rank value is normalized with respect to the total number of B tokens. Thus for each stop place confusion, the focus is on the rank distance of each CV stimulus from the mean of the target stop.

Confusion [ki] → [ti]

In the confusion study described in Chapter 4, [ki] stimuli were confused for alveolar place at 15%, whereas [ti] stimuli were rarely confused for velar place (2%). DT classifications predicted that NODE, TRI_ERR, BURST_NM, and marginally LIN_ERR and F2_00MS were the most informative features for discriminating between [ti] and [ki] in a pairwise discrimination task. Listener responses to [ki] stimuli were found to be significantly affected by their VOT value, surprisingly. No acoustic features correlated significantly with response rates to [ti] stimuli, perhaps due to their overall low rate of confusions.

The percentage overlap differences between [ki] and [ti] are high along all acoustic features (Figure 10). The lowest rates of overlap are for BURST_NM, LIN_ERR, and NODE. Features BURST_NM and F2_OOMS show dramatic asymmetries, with 100% of the [ti] tokens lying in the overlap range of values as compared to only 30% to 42% of the [ki] tokens. This indicates much more variation in values for the [ki] tokens along these two features. In cases of ambiguity, stimuli in the overlap region are hypothesized to be more likely to be confused for alveolar stop place than the reverse.

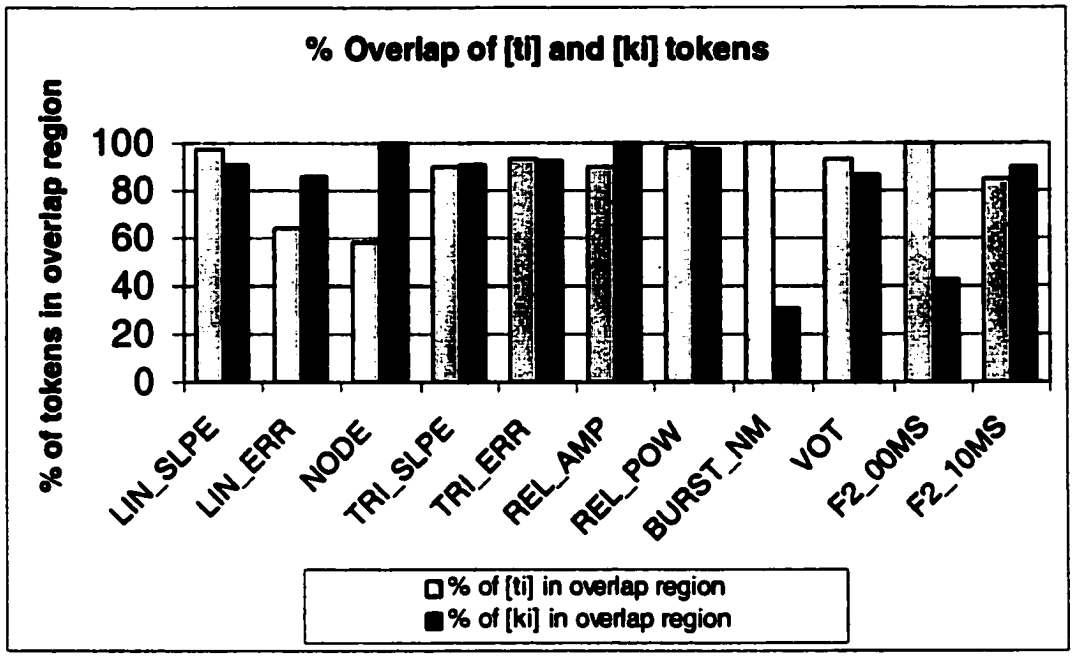


Figure 10. Percentage overlap of [ti] and [ki] by acoustic feature.

A multiple stepwise regression model was run on [ki] tokens that were confused for alveolar place by the rank values of all acoustic features available with respect to the [ti] mean for each feature. Only VOT rank normalized with respect to [t] was found to significantly correlate with the percentage of confusions of each stimuli (Beta = - 0.568, $p < 0.005$). [ki] stimuli with VOT values closest to the mean of [ti] VOT values were most likely to be confused for [ti] by listeners (Figure 11).

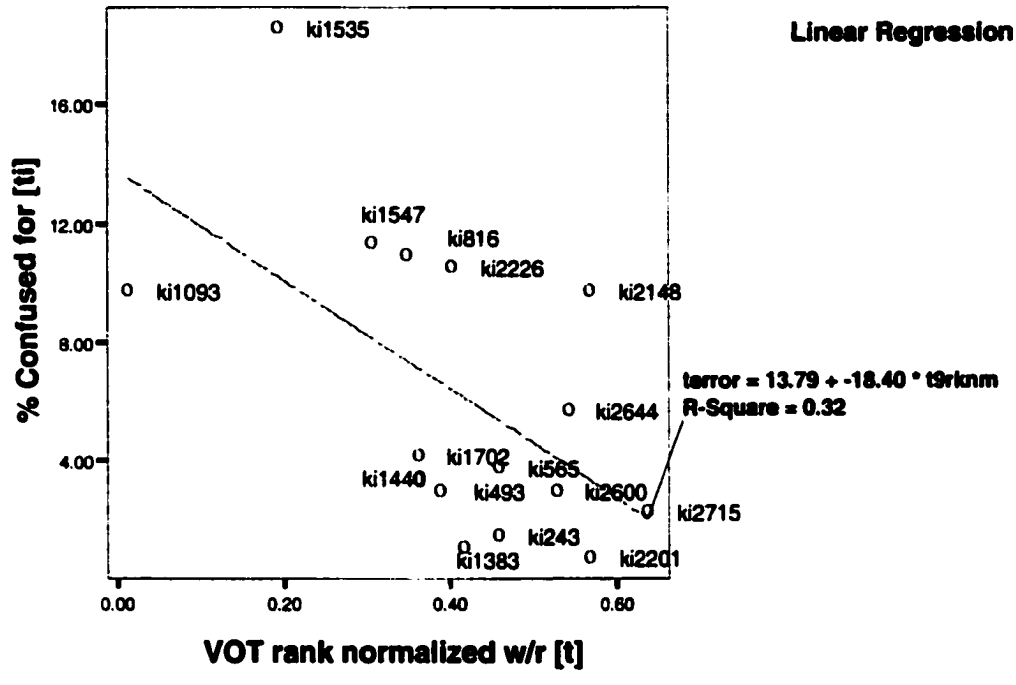


Figure 11. Linear regression of [ki] stimuli: The percentage of confusions for [ti] by their VOT value rank normalized with respect to the [ti] mean ($R^2 = 0.32$).

A similar multiple stepwise linear regression was run on the few [ti] tokens that were confused for [ki]. VOT and LIN_ERR when rank normalized with respect to the [ki] mean, were found to affect listener errors for velar place of articulation (Beta = 0.787, -0.346 respectively, $p < 0.005$) (Figure 12).

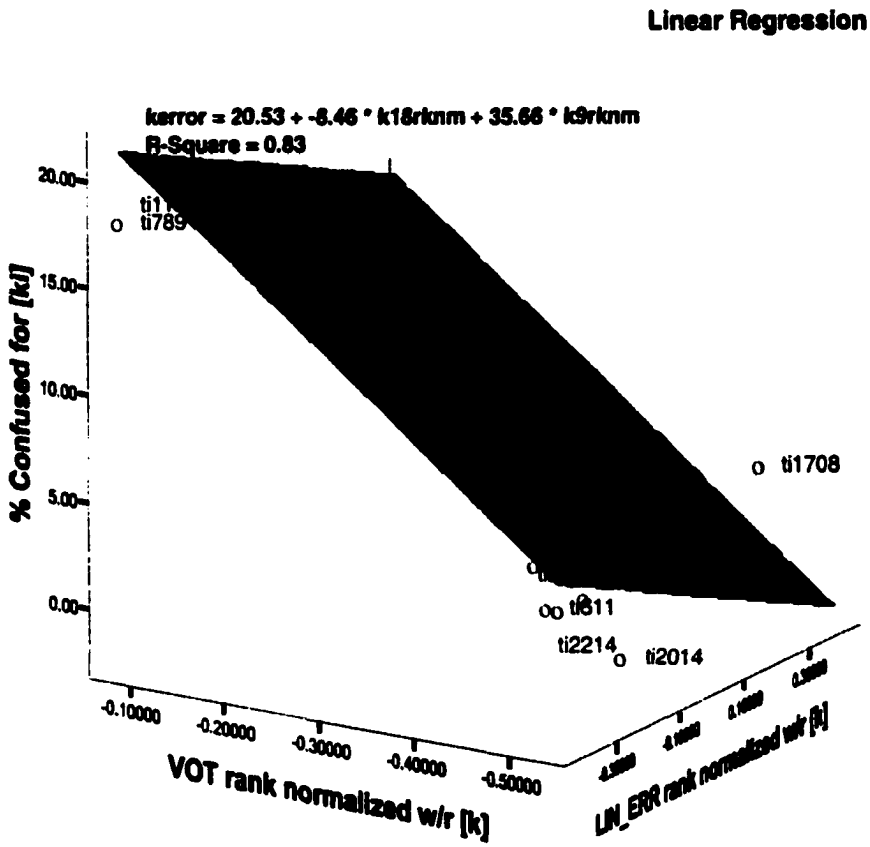


Figure 12. Linear regression of [ti] stimuli: The percentage of confusions for [ki] by their VOT and LIN_ERR values rank normalized with respect to the [ti] mean ($R^2 = 0.83$).

According to the non-parametric estimation of percentage of overlap, LIN_ERR is a feature along which [ki] and [ti] are asymmetric: a higher percentage of [ki] tokens (85%) lie in the overlap region than [ti] tokens (65%). This would suggest a preference for [ki] responses when the signal is ambiguous along this feature, which is confirmed by the significance of LIN_ERR as a factor in [ti] confusions for [ki]. This confusion was rarely attested in the experimental data, however (2.0%).

According to the non-parametric estimation of percentage of overlap, VOT is a feature along which [ki] and [ti] are symmetric, though with a high degree of overlap (85% to 95% of all [ki] and [ti] tokens lie in an overlap region). These results suggest that the asymmetric variation in stop production between [ki] and [ti] tokens is not the cause of asymmetries in listener confusions. In fact, both [ki] and [ti] are similarly distributed along the feature VOT, and VOT is found to be correlated both to [ki] → [ti] confusion rates as well as the relatively rare [ti] → [ki] confusion rates. The cause of the [ki] → [ti] confusion asymmetry does not appear to lie in differences in token-to-token variation.

Confusion [ku] → [tu]

The [ku] stimuli were confused for [tu] stimuli at 10.8%, whereas [tu] stimuli were rarely confused for velar place of articulation (1.1%). DT classifications found that the most useful features in [ku] / [tu] discrimination were NODE and LIN_SLPE alone or combined with TRI_ERR, based solely on the distributions of all tokens along these features. The rate of correct identification of [ku] stimuli was found to significantly correlate with their values of NODE, VOT, and F2_00MS. Although [tu] was rarely confused for other stop places, listeners' response rates for the correct identification of [tu] significantly correlated with VOT also.

The percentage of [tu] and [ku] tokens in the overlap region are most asymmetric for features NODE, REL_AMP, and F2_10MS. For all three of these features, a larger percentage of [tu] tokens lie in the overlap region than [ku] tokens (Figure 13). This suggests that [ku] tokens in the overlap region are more likely to be confused for alveolar

place than [tu] tokens in the overlap region would be confused for velar place. We will test the possible role of these features in stop place confusions using linear regression.

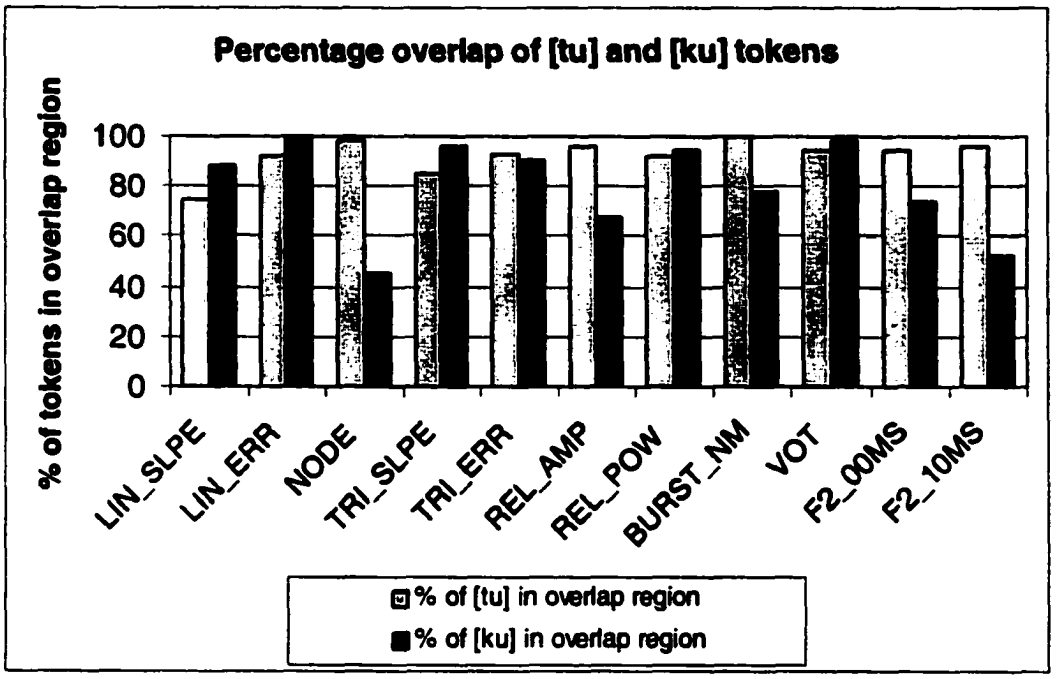


Figure 13. Percentage overlap for [tu] and [ku] tokens by feature.

A multiple stepwise linear regression was run on the [ku] stimuli that were confused for alveolar place in the perception study. The rank normalized values of F2_10MS and LIN_SLPE with respect to the [ku] mean along those values were found to correlate significantly with listener confusions for alveolar place (Beta = 0.737, 0.523, respectively, $p < 0.005$). [ku] tokens with LIN_SLPE values and F2_10MS values that approach the [tu] mean for those same features are more likely to be confused for [tu] by listeners (Figure 14).

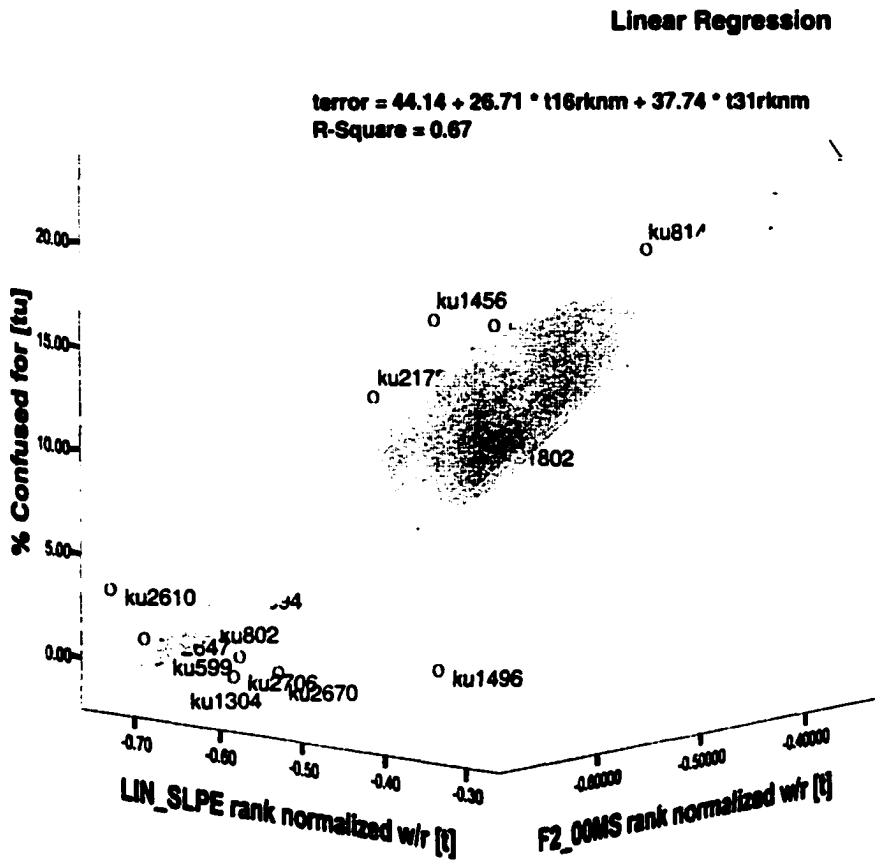


Figure 14. Linear regression of [ku] stimuli: the percentage of confusions for [tu] by their LIN_SLPE and F2_10MS values rank normalized with respect to the [tu] mean (R^2 = 0.67).

A multiple stepwise linear regression was also run on the percentage of [tu] tokens that were confused for velar place. Only the VOT rank normalized with respect to the [ku] mean significantly affected listener's rates of [tu] → [ku] confusions (Beta = 0.61, $p < 0.005$). However, upon closer examination, it was found that those [tu] tokens that were more likely to be confused for velar place are not closer to the [ku] mean VOT value; indeed, most [tu] tokens are centered around the mean value. Instead, they were

anomalous tokens that have VOT values much greater than the mean [ku] value. The percentage overlap method showed that a large percentage of both [tu] and [ku] tokens symmetrically overlap along VOT values. Thus, VOT was not predicted to trigger a asymmetric confusion between these two stops.

According to the non-parametric estimation of percentage of overlap, F2_10MS is a feature along which [ku] and [tu] show asymmetric degrees of overlap: 97% of all [tu] tokens but only 50% of all [ku] tokens lie in the overlap region. [ku] tokens with F2_10MS values that approach the [tu] mean were indeed more likely to be confused for [tu]. These results show that the asymmetric variation in stop production between [ku] and [tu] tokens may trigger asymmetries in listener confusions.

Confusion [pu] → [tu]

In the perceptual study, listeners confused [pu] stimuli for alveolar place 4.3% of the time, whereas [tu] stimuli were rarely confused for bilabial place (0.5%). DT classifications of [pu] and [tu] tokens showed that these tokens were differentiated most clearly along features REL_AMP, F2_10MS, F2_00MS, LIN_SLPE, and VOT. Correct responses in the perceptual study of [pu] tokens relied significantly on both F2_10MS and LIN_SLPE. [tu] tokens, however, were significantly more likely to be identified correctly when their VOT values were canonical (near the [tu] mean).

The most striking aspect of the non-parametric percentage of overlap between [pu] and [tu] tokens is the asymmetry in the amount of overlap of the features F2_00MS, F2_10MS, REL_AMP, and VOT. For these three features, the percentage of [pu] tokens

in the overlap range is less than the percentage of [tu] tokens in the overlap range (Figure 15).

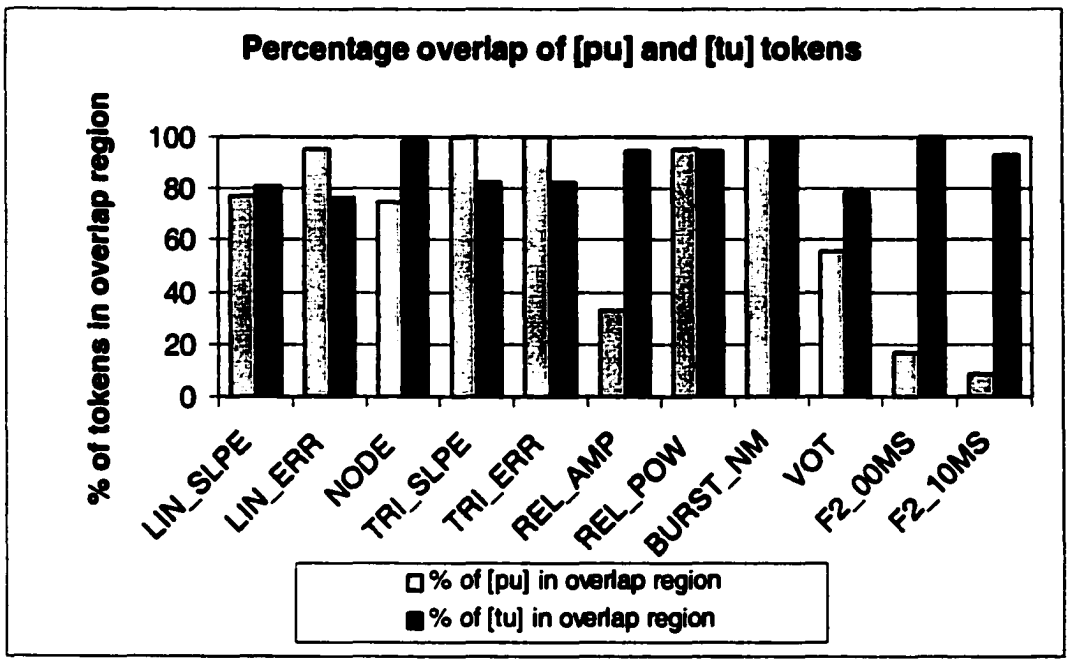


Figure 15. Percentage overlap of [pu] and [tu] tokens by feature.

A multiple stepwise linear regression was run on all acoustic features rank normalized with respect to the [tu] mean to determine their role, if any, in [pu] → [tu] confusions. A multiple stepwise linear regression was also run on all acoustic features rank normalized with respect to the [pu] mean to determine their role, if any, in the percentage of [tu] tokens confused for [pu] by listeners. No acoustic factor significantly affected listener rates for [pu] → [tu] confusions or for the relatively rare confusions from [tu] to bilabial place of articulation. The percentage overlap showed prominent asymmetries that were hypothesized to be responsible for increased rates of [pu] → [tu]

confusions with respect to [tu] → [pu] confusions, but listeners did not appear to respond according to these predictions. The cause of [pu] → [tu] asymmetries must lie elsewhere.

Confusion [pu] → [ku]

In the perceptual study, listeners confused [pu] stimuli for velar place 5.2% of the time, whereas [ku] stimuli were rarely confused for bilabial place (0.6%). DT classifications of [pu] and [ku] tokens showed that these tokens were differentiated most clearly along features LIN_ERR, REL_AMP, VOT, and TRI_ERR. Correct responses in the perceptual study of [pu] tokens relied significantly on both F2_10MS and LIN_SLPE. Unexpectedly, [ku] tokens were significantly more likely to be identified when their VOT, NODE, or F2_00MS values were canonical (near the [ku] mean).

[pu] and [ku] show low but conflicting overlap percentages along VOT and TRI_ERR (Figure 16). Formant features also appear to provide discriminatory information for this pair of stops. For both TRI_ERR and BURST_NM, a greater percentage of [pu] tokens lie in the overlap region than [ku] tokens. Thus, these two features are not predicted to play a role in [pu] → [ku] confusions. For VOT, LIN_SLPE, and the formant features, a greater percentage of [ku] tokens lie in the overlap region than [pu] tokens. These features are hypothesized to trigger an increase in [pu] → [ku] confusions for stimuli that are ambiguous along the features.

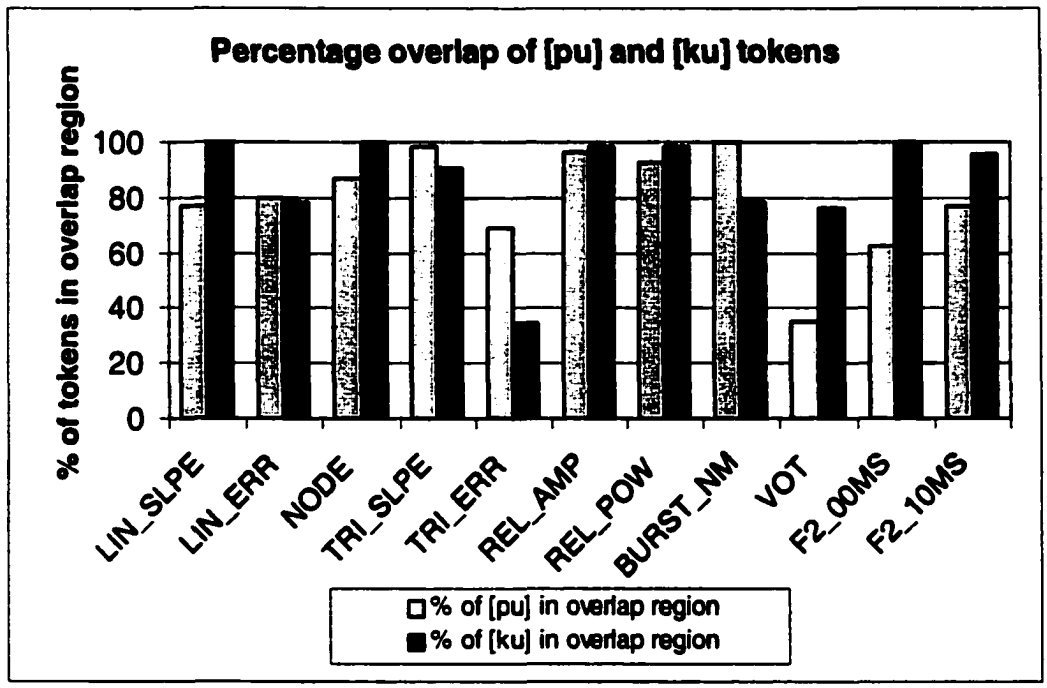


Figure 16. Percentage overlap for [ku] and [pu] by feature.

A multiple stepwise linear regression was first run on all acoustic features rank normalized with respect to the [ku] mean to determine their role, if any, in the percentage of [pu] tokens confused for [ku] by listeners. [pu] stimuli with F2_10MS approaching the [ku] mean along that feature, and especially those stimuli with values greater than the [ku] mean, were significantly more likely to be confused as [ku] tokens (Beta = 0.581, $p < 0.005$). [pu] and [ku] tokens in general are asymmetric in their distributions along F2_10MS. Also, F2_10MS was predicted to be a highly informative feature in the discrimination between [pu] and [ku] (Chapter 4). Listeners were shown to rely on F2_10MS for the identification of [pu] tokens. The asymmetry along this highly informative feature favors [ku] responses in [pu] and [ku] stimuli lying in the overlap

region. This would lead to higher rates of [pu] → [ku] confusions than [ku] → [pu] confusions.

A multiple stepwise linear regression showed that VOT, REL_AMP, and F2_10MS rank normalized with respect to the [pu] mean all combined to be significant factors in [pu]→[ku] confusions (Beta = -0.822, -0.284, 0.35, respectively, $p < 0.005$). The correlation by F2_10MS values, however, is contrary to expectations based on the percentage overlap of [pu] and [ku] tokens along this feature: confusions to bilabial place decrease as stimuli values approach the [pu] mean. A scatterplot revealed that one stimulus in particular was responsible for the majority of the 0.6% [ku] → [pu] confusions; its values emerge as significant.

Confusion [ka] ←→ [ta]

As a comparison with the asymmetric confusions analyzed in previous sections, the percentage overlap of the symmetric [ka] ←→ [ta] confusion is investigated. Listeners confused [ka] stimuli for alveolar place at 13.5% and [ta] stimuli for velar place at 7.6% of the time. DT classifications predicted that NODE and LIN_SLPE, possibly combined with TRI_ERR would be the most discriminatory features for [ka] and [ta] tokens, based solely on their distributions along these values. The rate of correct responses in the perceptual study for the identification of [ka] tokens correlated with VOT values. Listeners' rate of identification of [ta] tokens was significantly higher for tokens with mean values of VOT and LIN_SLPE.

[ta] and [ka] tokens show high percentages of overlap across all features extracted. No feature on its own is particularly informative, but recall that the DT

classification of [ta] and [ka] using all features had an accuracy of 90.6%, indicating that feature interactions in stop classification play a large role. No dramatic asymmetries between [ta] and [ka] were apparent at the feature level, though many small differences in overlap amount can be observed (Figure 17).

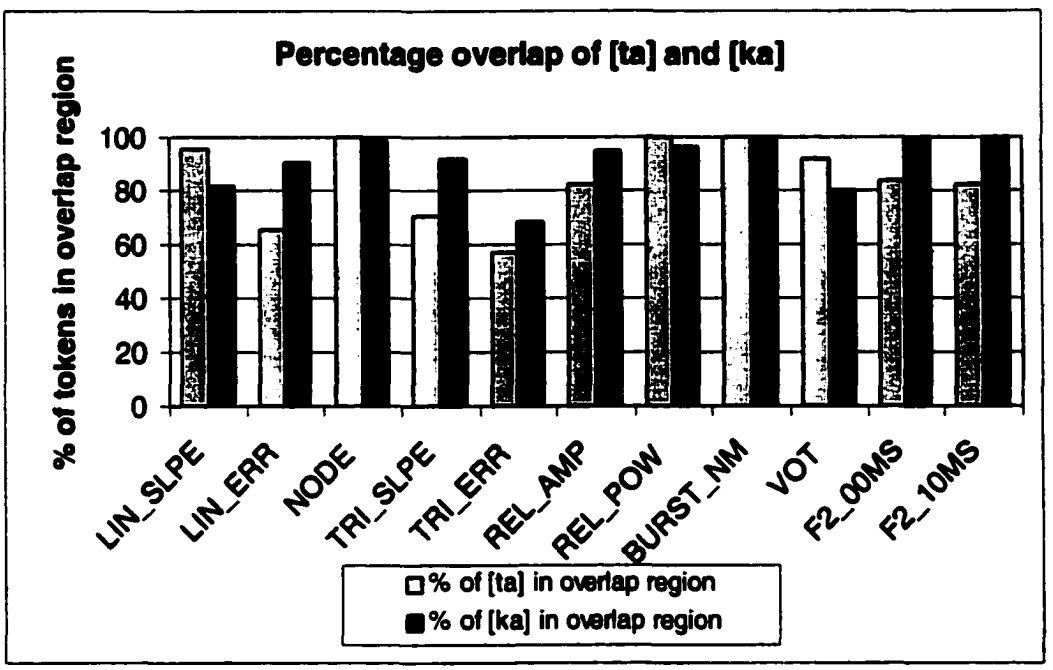


Figure 17. Percentage overlap for [ta] and [ka] tokens by feature.

A multiple stepwise linear regression was run on all acoustic features rank normalized with respect to the [ta] mean to determine their role, if any, in the percentage of [ka] tokens confused for [ta] by listeners. [ka] stimuli were significantly more likely to be confused for [ta] tokens when their VOT approached the [ta] mean (Beta = -0.765, $p < 0.005$). No other acoustic factor significantly affected listener confusion rates for alveolar place of articulation.

A multiple stepwise linear regression was also run on all acoustic features rank normalized with respect to the [ka] mean to determine their role if any in the percentage of [ta] tokens confused for velar place by listeners. [ta] tokens were significantly more likely to be confused for velar place of articulation when their REL_AMP value approached or was greater than the [ka] mean (Beta = .675, $p < 0.005$).

According to the non-parametric estimation of percentage of overlap, both VOT and REL_AMP are features along which [ka] and [ta] show slight opposing asymmetries in the amount of overlap. 81% of all [ta] tokens but 97% of all [ka] tokens lie in the overlap region along REL_AMP. This result predicts a preference for [ka] responses when stimuli that are ambiguous along REL_AMP are presented to listeners. Similarly, 95% of all [ta] tokens but only 80% of all [ka] tokens lie in the overlap region along VOT. This predicts that listeners presented with stimuli that are ambiguous along VOT will show a preference for [ta] responses. These results overall show that asymmetries between [ka] and [ta] tokens at the feature level affect listener confusions. In this case, however, the resulting confusions are symmetric at the categorical level, perhaps because [ta] responses are favored for stimuli in the overlap region of VOT, but [ka] responses are favored for stimuli in the overlap region of REL_AMP.

5.3 Summary and discussion

Results from correlations of listener rates of errors in the perceptual study and percentage overlap of all CV tokens along relevant features are mixed. Results from [ki]→ [ti] do not support the hypothesis that listeners favor stop places associated with less token-to-token variation in ambiguous cases. Instead, the rank distance of stimuli from VOT means is found to influence error rates for both [ki] to [ti] and [ti] to [ki] confusions. In addition, rank distance of stimuli from LIN_ERR is found to favor [ki] responses, the rarely attested confusion.

[pu] → [tu] also failed to support the hypothesis that listeners favor stop places associated with less token-to-token variation in ambiguous cases. Features F2_00MS, F2_10MS, REL_AMP, and VOT were all good candidates for favoring [tu] responses when listeners were presented with ambiguous stimuli, since the percentage of [tu] tokens in the overlap region is much higher than the percentage of [pu] tokens along all four features. Listeners in the perception study did not appear to respond according to these predictions, however.

Support for the hypothesis was found in the asymmetric [ku] → [tu] confusion. [tu] responses were predicted to be greatly favored in the overlap region along F2_10MS, a feature which is known to be used by listeners for [ku] identification, due to the relative percentage of overlap of [ku] and [tu] tokens along this feature (46% and 97%, respectively), and F2_10MS was indeed found to be significant in the percentage of listener errors from [ku] to [tu] in the perception study.

Similarly, F2_10MS was predicted to favor [pu] → [ku] confusions, since there is a greater percentage of overlap of [ku] tokens than [pu] tokens along this feature (97% and 80% overlap, respectively). F2_10MS is also known to be an important cue in [pu] / [ku] distinctions, as listeners relied on second formant onset for both [pu] and [ku] identification (Chapter 4). A multiple linear regression showed that [pu] stimuli with F2_10MS values approaching the [ku] mean for that feature, and especially those stimuli with values greater than the [ku] mean, were significantly more likely to be confused as [ku] tokens.

Results from the symmetric confusion [ka] ←→ [ta] also supported the hypothesis that listeners use their knowledge of token distributions along relevant features to make categorical stop decisions. The percentages of overlap showed the fewest asymmetries for all stop pairs analyzed. REL_AMP was predicted to slightly favor [ka] responses and VOT was predicted to slightly favor [ta] responses. VOT was used by listeners in both [ka] and [ta] identification (Chapter 4), contrary to DT predictions, and was also found to be significant in listeners' confusions of [ka] → [ta], the direction with the greatest error rates (13.5%). [ta] → [ka] errors (7.6%), however, were significantly correlated to the rank distance of stimuli along REL_AMP. The combination of the two leads to the symmetric stop place confusion [ka] ←→ [ta].

6 Conclusion

Results from the perception study presented in Chapter 4 were used to evaluate several explanations for the asymmetry of certain stop place confusions. The claim that affinities between vowels and consonant account for response biases toward alveolars preceding high front contexts and bilabials preceding back rounded vowels was not supported, since error rates did not significantly vary across the two test versions. In one version of the test, the identity of the vowel varied throughout the test blocks, while in the other, each test block presented CV stimuli with a single vowel. Additionally, the claim offered no explanation for the most prominent asymmetric confusions found in the current experiment, [k] → [t] in all vocalic contexts.

The current perceptual experiment supports the claim that asymmetries in stop consonant confusions are tied to the frequency of the phoneme in the language (Lyublinskaya 1966, Vitevitch et al. 1999). The frequency of word tokens containing a given CV segment significantly correlates with the percentage of incorrect responses for that CV in the perception experiment for both the Switchboard and Brown corpora. CVs that are more frequent in a listener's lexicon are more likely to be given as incorrect responses in the stop place identification task. [ta] stimuli appeared to be an exception to this overall trend, since they were commonly confused in the perception study but relatively infrequent in the two corpora. [ka] → [ta] confusions are most likely triggered by distributional differences between the two stops along formant onset features.

Although markedness (due to phonological considerations, frequency effects, or production constraints) appears to play a role in asymmetric stop consonant confusions,

acoustic causes of confusions are shown to be primary by Chang et al. (2001), at least in the case of [ki] → [ti] confusions. The high front vowel is thought to neutralize formant onset cues to stop place, leaving only the relatively non-robust spectral cues of the bursts. Velar and alveolar stops have similar burst spectra in the context of high front vowels, with the exception that velars have an extra mid-frequency peak. The mid-frequency cue is often degraded due to entropy, and in the case of [ki], likely to be confused for an alveolar stop place. Other asymmetric confusions may also be triggered by an extra feature, as in the case of [ki] → [ti], but those features must be investigated case by case. Results from the perception study that indicate the features that correlate with specific listener errors (Section 5) point to regions of the acoustic signal that should be further investigated for possible causes of asymmetric confusions.

The final hypothesis of the causes of asymmetric confusions was tested by examining asymmetries in the percentage of overlap between pair of stops. Results from correlations of listener rates of errors in the perceptual study and percentage overlap of all CV tokens along relevant features were mixed. Asymmetries between the distribution of a pair of stops along a feature appear to lead to asymmetries in stop place confusions only when the feature is also known to be used by listeners in stop place identification, as it was for [ku] → [tu] and [pu] → [tu]. Mini-sound changes such as these may lead to historical sound changes, given the right sociolinguistic environment. Additional evidence was found in the one symmetric confusion examined, [ka] ←→ [ta], for which two competing features were found to be significant in listener errors in either direction.

Chapter 6

Conclusion

This thesis is a first attempt at investigating the causes of asymmetric stop place confusions and at challenging explanations for this phenomenon that rely solely on the inherently symmetric notion of *similarity*. Although some accuracy and faithfulness to the human auditory processing system was sacrificed by the use of a large set of semi-automatically extracted cues, the combination of decision trees (DTs) and classic perceptual experimentation revealed the following:

- *Frequent confusions of bilabial and velar stop place for alveolars may be due to their relative distributions in the multidimensional acoustic-perceptual space.* Bilabials and velars are acoustically similar to alveolars in many contexts, but are distinct from one another, especially along features such as the second formant onset, VOT, and relative amplitude.

- *The relative roles of stop place cues vary by stop place and by context.* VOT and the gross spectral shape of the burst offer a three-way classification between bilabials, alveolars, and velars. Relative amplitude and formant onsets are most useful for distinguishing bilabials from other stop places. Multiple bursts are useful for indicating velar place. F2 onset is a primary feature for distinguishing

velars from other stops in both [a] and [u] contexts, but is secondary to burst characteristics in the [i] context.

- *Listener performance is affected by the amount of discriminatory information in the acoustic signal.* For example, listener errors in the context of [i] are partially due to a relatively impoverished signal in this environment; there are fewer discriminatory cues for stop place available to the listener. An additional result is that the method developed to estimate the amount of discriminatory information in the acoustic signal (accuracy of DT classifiers of stop place categories) can predict listener performance with some accuracy.

- *Phonotactic frequency in the lexicon, degradation of non-robust cues, and differences in production variation all affect perceptual error rates differentially by CV context.* The results show that humans employ whatever cues (phonetic or non-phonetic) are available for the categorization of speech sounds, especially when primary acoustic cues are ambiguous. The same factors that cause confusions in unaltered CV tokens are likely to be the cause for parallel historical sound changes.

The methodology adopted in the current project offers a new direction in perceptual studies, one that relies on the same large set of unaltered tokens for both the acoustic analysis and the stimuli set of the perception study. The larger set of tokens and acoustic features lends itself to machine learning techniques for mapping out phonemic

categories within a multidimensional acoustic-perceptual space. This approach transfers easily to other sets of phonemes as well as to other languages. Recent research in speech recognition has shown the approach to phonetic classification developed here to be useful for data-driven techniques to temporal front-end adaptation (Sönmez et al. 2000).

Improvements could be made in the DSP techniques used to extract the acoustic properties from the signal such that dynamic cues, such as formant transitions and changes in the spectrum following the burst release, could be included in the set of candidate cues. Training DTs on improved auditory models that are more faithful to the human internal representation of speech sounds (Ghitza 1993; Ghitza & Sondhi 1997) would result in a more accurate relative ranking of acoustic features with respect to human judgments of stop place. Determining precisely how listeners integrate the information provided by the large and variable set of acoustic cues available for the purpose of phoneme categorization is an interesting problem for years to come.

References

Ahmed, R. and S. S. Agrawal. 1969. Significant features in the perception of (Hindi) consonants. *Journal of the Acoustical Society of America* 45, 3: 758-763.

Andersen, H. 1973. Abductive and deductive change. *Language* 49: 765-793.

Bacchiani, M. and M. Ostendorf. 1998. Using automatically-derived acoustic subword units in large vocabulary speech recognition. *Proceedings of the International Conference on Spoken Language Processing* 5: 1843-1846.

Bell, T. S., D. D. Dirks, and E. C. Carterette. 1989. Interactive factors in consonant confusion patterns. *Journal of the Acoustical Society of America* 85, 1: 339-346.

Bennett, W. H. 1969. Pre-Germanic /p/ for Indo-European /k^w/. *Language* 45: 243-7.

Blumstein, S. E. 1986. On acoustic invariance in speech. In J. Perkell and D. Klatt (Eds.) *Invariance and Variability in Speech Processes*. Lawrence Erlbaum: Hillsdale, NJ. 178-201.

Blumstein, S. E. and K. N. Stevens. 1979. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America* 66, 4: 1001-1017.

Blumstein, S. E. and K. N. Stevens. 1981. Phonetic features and acoustic invariance in speech. *Cognition* 10: 25-32.

Bonebrake, V. 1979. Historical labial-velar changes in Germanic. *Acta Universitatis Umensis (Umeå Studies in the Humanities)* 29, Sweden.

Bonneau, A., L. Djezzar, and Y. Laprie. 1996. Perception of the place of articulation of French stop bursts. *Journal of the Acoustical Society of America* 100, 1: 555-564.

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and Regression Trees*. Wadsworth: Belmont, CA.

Buntine, W. and R. Caruana. 1992. *Introduction to IND version 2.1 and recursive partitioning*. NASA Ames Research Center. December 31.

Chafe, W. L. 1964. Another look at Siouan and Iroquoian. *American Anthropologist* 66: 852-62.

Chang, S. S. 1999. Vowel dependent VOT variation. *Proceedings of the XIVth International Congress of Phonetic Sciences* Volume 2: 1021. San Francisco, CA.

Chang, S. S., M. C. Plauché, and J. J. Ohala. 2001. Markedness and consonant confusion asymmetries. In K. Johnson and E. Hume (Eds.) *The Role of Perceptual Phenomena in Phonological Theory*. 79-101. Academic Press: San Diego, CA.

Chen, M. 1973. *Nasals and Nasalization in the History of Chinese.* Doctoral dissertation, University of California, Berkeley.

Chomsky, N. and M. Halle. 1968. *Sound Patterns of English.* Harper and Row: New York, NY.

Cole, R., R. Stern, and M. Lasry. 1986. Performing fine phonetic distinctions: Templates versus features. In J. S. Perkell and D. M. Klatt (Eds). *Variability and Invariance in Speech Processes*. Lawrence Erlbaum: Hillsdale, NJ.

Cooper, F. S., P. C. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman. 1952. Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America* 24, 6: 597-606.

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

Delgutte, B. 1980. Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. *Journal of the Acoustical Society of America* 68, 3: 843-857.

Delogu, C., A. Paoloni, P. Ridolfi, and K. Vagges. 1995. Intelligibility of speech produced by text-to-speech systems in good and telephonic conditions. *Acta Acustica* 3: 89-96.

Dorman, M. F., M. Studdert-Kennedy, and L. J. Raphael. 1977. Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception & Psychophysics* 22, 2: 109-122.

Dorman, M. F. and P. C. Loizou. 1996. Relative spectral change and formant transitions as cues to labial and alveolar place of articulation. *Journal of the Acoustical Society of America* 100, 6: 3825-3830.

Fant, G. 1973. *Speech Sounds and Features*. MIT Press: Cambridge, MA.

Fischer-Jørgensen, E. 1967. Acoustic analysis of stop consonants. *Readings in Acoustic Phonetics*. In Ilse Lehiste (Ed.), Reprinted from: *Miscellenea Phonetica*, Volume II (1954). MIT Press: Cambridge, MA.

Fowler, C. A. 1994. Invariants, specifier, cues: An investigation of locus equations as information for place of articulation. *Perception and Psychophysics* 55: 597-610.

Kucera, H. and W. N. Francis. 1967. *Computational Analysis of Present-day American English*. Brown University Press: Providence, RI.

Fujimura, O. 1961. Bilabial stop and nasal consonants: A motion picture study and its acoustical implications. *Journal of Speech and Hearing Research* 4: 233-246.

Garner, W. R. 1978. Aspects of a stimulus: Features, dimensions, and configurations. In E. Rosch and B. B. Lloyd (Eds.) *Cognition and Categorization.* Lawrence Erlbaum: Hillsdale, NJ. 99-133.

Ghitza, O. 1993. Adequacy of auditory models to predict human internal representation of speech sounds. *Journal of the Acoustical Society of America* 93, 4: 2160-2171.

Ghitza, O. and M. M. Sondhi. 1997. On the perceptual distance between speech segments. *Journal of the Acoustical Society of America* 101, 1: 522-529.

Gilmore, G. C., H. Hersh, A. Caramazza, and J. Griffin. 1979. Multidimensional letter similarity derived from recognition errors. *Perception & Psychophysics* 25, 5: 425-431.

Glass, J., J. Chang, and M. McCandless. 1996. A probabilistic framework for feature-based speech recognition. *Proceedings of the International Conference on Spoken Language Processing,* 2277-2280. Philadelphia, PA.

Godfrey, J., E. Holliman, and J. Mc Daniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development, *International Conference on Acoustics, Speech, and Signal Processing* 1, 517-520.

Gold, B. and N. Morgan. 2000. *Speech and Audio Signal Processing.* John Wiley: New York, NY.

Grierson, G.A. 1906. (2nd edition 1969). *The Píśāca Languages of North-Western India.* Munishiram Manoharlal, Delhi.

Guion, S. G. 1996. *Velar Palatalization: Coarticulation, Perception and Sound Change.* Doctoral dissertation. University of Texas, Austin.

Guion, S. G. 1998. The role of perception in the sound change of velar palatalization. *Phonetica* 55: 18-52.

Guthrie, M. 1967-1970. *Comparative Bantu*. Gregg. 4 Volumes.

Haas, M. R. 1969. *The Prehistory of Language*. Mouton: The Hague.

Hedrick, M. S. and R. N. Ohde. 1993. Effect of relative amplitude of frication on perception of place of articulation. *Journal of the Acoustical Society of America* 94, 4: 2005-2026.

Hedrick, M. and W. Jesteadt. 1996. Effect of relative amplitude, presentation level, and vowel duration on perception of voiceless stop consonants by normal and hearing-impaired listeners. *Journal of the Acoustical Society of America* 100, 5: 3398-3407.

Hock, H. H. 1938 (2nd edition 1991). *Principles of Historical Linguistics*. Mouton de Gruyter: Berlin.

Hoole, P., K. Munhall, and C. Mooshammer. 1998. Do airstream mechanisms influence tongue movement paths? *Phonetica* 55: 131-146.

Homburger, L. 1949. *The Negro-African Languages*. Routledge & Kegan Paul: London.

Houde, R. A. 1968. A study of tongue body motion during selected speech sounds. *Speech Communication Research Laboratory*. Monograph 2. UC Santa Barbara: Santa Barbara, CA.

Jakobson R., G. Fant, and M. Halle. 1952.*Preliminaries to speech analysis*. MIT Press: Cambridge, MA.

Keating, P. and A. Lahiri. 1993. Fronted velars, palatalized velars, and palatals. *Phonetica* 50: 73-101.

Kewley-Port, D. 1982. Measurement of formant transitions in naturally produced stop consonant-vowel syllables. *Journal of the Acoustical Society of America* 72, 2: 379-389.

Kewley-Port, D. 1983. Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America* 73, 1: 322-335.

Kingston, J. 1992. The phonetics and phonology of perceptually motivated articulatory covariation. *Language and Speech* 35, 1-2: 99-113.

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

Kuhl, P. 1992. Human adults and human infants show a perceptual magnet effect for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics* 50, 2: 93-107.

Lahiri, A., L. Gewirth, and S. E. Blumstein. 1984. A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study. *Journal of the Acoustical Society of America* 76, 2: 391-404.

Lathrop, T.A. 1980. The evolution of Spanish: An introductory historical grammar. *Cuesta-Hispanic Monographs.* Newark, DE.

Lehiste, I. 1970. *Suprasegmentals.* MIT Press: Cambridge, MA.

Lehiste, I. and G. Peterson. 1959. Vowel amplitude and phonemic stress in American English. *Journal of the Acoustical Society of America* 31, 4: 428-435.

Li, F. K. 1977. *A Handbook of Comparative Tai.* University Press of Hawaii: Manoa, HI.

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

Liberman, A. M. & Mattingly, I. G. 1985. The motor theory of speech perception revised. *Cognition* 21: 1-36.

Lindblom, B. 1986. On the origin and purpose of discreteness and invariance in sound patterns. In J. Perkell and D. Klatt (Eds.) *Invariance and Variability in Speech Processes*. Hillsdale, NJ: Lawrence Erlbaum. 493-510.

Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H&H theory. In W. Hardcastle and A Marchai (Eds.) *Speech Production and Speech Modeling*. Kluwer: Dordrecht. 403-439.

Lindblom, B., S. Guion, S. Hura, S.-J. Moon, and R. Willerman. 1995. Is sound change adaptive? *Rivista di Linguistica* 7, 1: 5-37.

Lisker, A. 1964. A cross language study of voicing in initial stops; acoustic measurements. *Word* 20, 3: 384-422.

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

Longacre, R. 1967. Systematic comparison and reconstruction. *Handbook of Middle American Indians* 5:117-59.

Lyublinskaya, V. V. 1966. Recognition of articulation cues in stop consonants in transition from vowel to consonant. *Soviet Physics-Acoustics* 12, 2: 185-192.

Maddieson, I. 1984. *Patterns of Sounds*. Cambridge University Press: Cambridge, UK.

Maddieson, I. 1986. The size and structure of phonological inventories: Analysis of UPSID. In John J. Ohala and Jeri Jaeger (Eds.) *Experimental Phonology*. Academic Press: Orlando, FL. 105-123.

Maddieson, I. 1997. Phonetic Universals. In W. J. Hardcastle and J. Laver (Eds.) *The Handbook of Phonetic Sciences*. Blackwell Publishers: MA.

Malkiel, Y. 1963. The interlocking of narrow sound change, broad phonological pattern, level of transmission, areal configuration sound symbolism. *Archivum Linguisticum* 15, 2: 144-173.

Meillet, A. and Vendryes, J. 1924. *Traité de Grammaire Comparée des Langues Classiques*. Libraire Ancienne Edouard Champion: Paris.

Miller, G. A. and P. E. Nicely. 1955. An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America* 27, 2: 338-352.

Miller, M.I. and M. B. Sachs. 1983. Representation of stop consonants in the discharge patterns of auditory-nerve fibers. *Journal of the Acoustical Society of America* 74, 2: 502-517.

Mullennix, J. W. and D. B. Pisoni. 1990. Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics* 47, 4: 379-390.

Neagu, A. 1998. *Represéntations phonétiques et identification des syllables occlusive-voyelle en français*. Doctoral dissertation. Institut National Polytechnique, Grenoble.

Nearey, T. M. & B. L. Rochet 1994. Effects of place of articulation and vowel context on VOT production and perception for French and English stops. *Journal of the International Phonetic Association* 24, 1:1-18.

Nieuwint, P. J. G. M. 1981. What happened to middle English /(u)(x)/? *Neophilologus* 65: 440-467.

Ohala, J. J. 1971. Monitoring soft palate activity in speech. *Project on Linguistic Analysis (Berkeley, CA)* 13: J01-J015.

Ohala, J. J. 1974. Experimental historical phonology. In J. M. Anderson and C. Jones (Eds.) *Historical Linguistics II. Theory and Description in Phonology.* North Holland Publishing Company: Amsterdam. 353-389.

Ohala, J. J. 1978. Southern Bantu vs. the world: The case of palatalization of labials. *Proceedings of the Berkeley Linguistic Society* 4: 370-386. Berkeley Linguistics Society: Berkeley, CA.

Ohala, J. J. 1981. The listener as a source of sound change. In R. A. Hendrick, C. S. Masek, and M. F. Miller (Eds.) *Proceedings of the Chicago Linguistics Society* 178-203. Chicago Linguistics Society: Chicago, IL.

Ohala, J. J. 1983. The phonological end justifies any means. In S. Hattori and K. Inoue (Eds.) *Proceedings of the XIIIth International Congress of Linguists.* Tokyo, Japan. 232-243.

Ohala, J. J. 1985. Linguistics and automatic processing of speech. In R. De Mori and C. Y. Suen (Eds.) *New Systems and Architectures for Automatic Speech Recognition and Synthesis.* Springer-Verlag: Berlin. 447-475.

Ohala, J. J. 1990. The phonetics and phonology of aspects of assimilation. In J. Kingston and M. Beckman (Eds.) *Between the Grammar and the Physics of Speech.* Cambridge University Press: Cambridge, UK. 258-275.

Ohala, J. J. 1993. Sound change as nature's speech perception experiment. *Speech Communication* 13: 155-161.

Ohala, J.J. and J. Lorentz. 1977. The story of [w]: an exercise in the phonetic explanation for sound patterns. *Proceedings of the Berkeley Linguistic Society* 3: 577-599. Berkeley Linguistics Society: Berkeley, CA.

Ohala, M. 1995. Acoustic study of VC transitions for Hindi stops. *Proceedings of the XIIth International Congress of Phonetic Sciences* 4: 22-25. Stockholm.

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

Ohde, R. N. and K. N. Stevens. 1983. Effect of burst amplitude on the perception of stop consonant place of articulation. *Journal of the Acoustical Society of America* 74, 3: 706-714.

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

Pagliuca, W. 1982. *Prolegomena to a Theory of Articulatory Evolution.* Doctoral dissertation. State University of New York, Buffalo.

Perkell, J. 1969. *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study.* MIT Press: Cambridge, MA.

Plauché, M. C., C. Delogu, and J. J. Ohala. 1997. Asymmetries in consonant confusion. *Proceedings of Eurospeech 1997.* 2187-2190. Rhodes, Greece.

Plauché, M. and K. Sönmez. 2000. Machine learning techniques for the identification of cues for stop place. *Proceedings of the 6th International Conference on Spoken Language Processing*, Vol. 1: 548-551. Beijing.

Repp, B. H. 1984a. The role of release bursts in the perception of [s]-stop clusters. *Journal of the Acoustical Society of America* 75, 4: 1219-1230.

Repp, B. H. 1984b. Closure duration and release burst amplitude cues to stop consonant manner and place of articulation. *Language and Speech* 27, 3: 245-254.

Repp, B. H. and Lin, H.-B. 1989. Acoustic properties and perception of stop consonant release transients. *Journal of the Acoustical Society of America* 85, 1: 379-395.

Riordan, C. J. 1977. Control of vocal-tract length in speech. *Journal of the Acoustical Society of America* 62, 4: 998-1002.

Russell, S. and P. Norvig. 1995. *Artificial Intelligence, A Modern Approach*. Prentice Hall: Englewood Cliffs, NJ.

Schatz, C. D. 1954. The role of context in the perception of stops. *Language* 30, 1: 47-56.

Shriberg, E. E. 1992. Perceptual restoration of filtered vowels with added noise. *Language and Speech* 35, 1-2: 127-136.

Singh, S. and J. W. Black. 1966. Study of twenty-six intervocalic consonants as spoken and recognized by four language groups. *Journal of the Acoustical Society of America* 39: 372-387.

Smits, R. 1995. *Detailed Versus Gross Spectro-Temporal Cues for the Perception of Stop Consonants*. Doctoral dissertation. Technische Universiteit, Eindhoven.

Smits, R. 2000. Temporal distribution of information for human consonant recognition in VCV utterances. *Journal of Phonetics* 27:111-135.

Sönmez, K., M. Plauché, E. Shriberg, and H. Franco. 2000. Consonant discrimination in elicited and spontaneous speech: A case for signal-adaptive front ends in ASR. *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing. Vol.1: 325-329.

Stevens, K. N. 1960. Spectra of fricative noise in human speech. *Language and Speech* 3: 32-49.

Stevens, K. N. 1999. *Acoustic Phonetics*. MIT Press: Cambridge, MA.

Stevens, K. N. and D. H. Klatt. 1974. Role of formant transitions in the voiced-voiceless distinction for stops. *Journal of the Acoustical Society of America* 55, 3: 653-659.

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

Summerfield, Q. and M. Haggard. 1977. On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America* 62, 2: 435-448.

Sussman, H. M., H. A. McCaffrey, and S. Matthews. 1991. An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America* 90, 3: 1309-1325.

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

Vitevitch, M. S. and P. A. Luce. 1999. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language* 40: 374-408.

Vitz, P. C. and B. S. Winkler. 1973. Predicting the judged "similarity of sound" of English words. *Journal of Verbal Learning and Verbal Behavior* 12: 373-388.

Wang, M. D. and R. C. Bilger. 1973. Consonant confusions in noise: A study of perceptual features. *Journal of the Acoustical Society of America* 54, 5: 1248-1266.

Weinstein, C., S. McCandless, L. Mondshein, and V. Zue. 1975. A system for acoustic-phonetic analysis of continuous speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 23: 54-67.

Weismer, G. 1980. Control of the voicing distinction for intervocalic stops and fricatives: some data and theoretical considerations. *Journal of Phonetics* 8: 427-438.

Wickelgren, W. 1966. Distinctive features and errors in short-term memory for English consonants. *Journal of the Acoustical Society of America* 39: 388-398.

Wickelgren, W. A 1976. Phonetic Coding and Serial Order. In E. Carterette and M. Friedman (Eds.) *Handbook of Perception Vol 7: Language and Speech.* Academic Press, NY.

Winitz, H., M. E. Scheib, and J. A. Reeds. 1972. Identification of stops and vowels for the burst portion of /p, t, k/ isolated from conversational speech. *Journal of the Acoustical Society of America* 51, 4(2): 1309-1317.

Zue, V. W. 1976. *Acoustic Characteristics of Stop Consonants: A Controlled Study.* Lincoln Lab: Lexington, MA.