The stop-like modification of /ð/: A case study in the analysis and handling of speech variation

by

Sherry Y. Zhao

B.S., E.E.C.S., Massachusetts Institute of Technology 2002

M.Eng., E.E.C.S., Massachusetts Institute of Technology 2003

Submitted to the
Harvard-MIT Division of Health Sciences and Technology
In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Speech and Hearing Bioscience and Technology

at the

Massachusetts Institute of Technology

September 2007

Author ______________________________
Department of Health Sciences and Technology
July 30, 2007

Certified by ______________________________
Kenneth N. Stevens
Clarence J. LeBel Professor of Electrical Engineering
Thesis Supervisor

Accepted by ______________________________
Martha L. Gray
Edwin Hood Taplin Professor of Medical and Electrical Engineering
Co-Director, Harvard-MIT Health Sciences and Technology

Page intentionally left blank

Acknowledgements

I thank my advisor Prof. Ken Stevens for sparking my interests in the field of speech communication and fostering me as a research scientist. It is truly an honor to be his student. I thank my thesis committee chair, Dr. Stefanie Shattuck-Hufnagel, for her insightful comments and constant encouragement. Her enthusiasm in my research inspires me to always do my best. I thank Prof. Michael Kenstowicz for being on my graduate committees as well as always taking the time to read and discuss my research. I truly appreciate the linguistics perspective that he brings to my thesis. I thank Dr. Janet Slifka for taking the time to be on all of my graduate committees. Her feedback has been crucial in the development and implementation of my research ideas.

Furthermore, I thank my academic advisors Dr. Andrew Oxenham and Dr. Jennifer Melcher for helping me throughout the HST Speech and Hearing Program. I thank Dr. Jim Glass for being on my proposal advisory committee and encouraging me to pursue the topic of this thesis.

I thank Xuemin Chi, Steven Lulich, Daryush Mehta, Xiaomin Mou, Tony Okobi, Yoko Saikachi, and Tara Sainath for lively conversations about life and work. Their friendship has made my experience rich and satisfying. I thank Arlene Wint for always being there to help me with lab and work-related details. She is indeed the best lab mother! Also, I thank everyone at the Speech Communication Group for the positive learning environment they help to create. I could not have worked at a more wonderful lab. I also thank everyone I have worked with and befriended in the HST Speech and Hearing Program. There are truly too many people to thank.

Last but certainly not least, I thank Mom, Dad, and my husband Hanching. Their support and encouragement have carried me through the ups and downs of graduate school. Without them, none of this would have been possible.

This work is supported by NIH Grants DC02978 and T32DC00038.

Page intentionally left blank

The stop-like modification of /ð/: A case study in the analysis and handling of speech variation

by

Sherry Y. Zhao

Submitted to the Harvard-MIT Division of Health Sciences and Technology on July 30, 2007 In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Abstract

Phonetic variation is pervasive in everyday speech. Studying these variations is essential for building acoustic models and lexical representations that effectively capture the variability of speech. This thesis examines one of the commonly-occurring phonetic variations in English: the stop-like modification of the dental fricative /ð/. This variant exhibits a drastic change from the canonical /ð/; the manner of production is changed from one that is fricative to one that is stop-like. Furthermore, the place of articulation of stop-like /ð/ has been a point of uncertainty, leading to the confusion between stop-like /ð/ and /d/. This thesis aims to uncover the segmental context of stop-like /ð/, possible causes of the modification, whether the dental place of articulation is preserved despite modification, and if there are salient acoustic cues that distinguish between stop-like /ð/ and /d/.

Word-initial /ð/ in the read speech of the TIMIT Database, the task-oriented spontaneous speech of the AEMT Corpus, and the non-task-oriented spontaneous speech of the Buckeye Corpus are examined acoustically. It is found that stop-like /ð/ occurs most often when it is preceded by silence or when preceded by a stop consonant. The occurrence is less frequent when /ð/ is preceded by a fricative or an affricate consonant. This modification rarely occurs when /ð/ is preceded by a vowel or liquid consonant. The findings suggest that possible factors that may contribute to the stop-like modification of /ð/ include physiological mechanisms of speech production, prosody, and/or other aspects of speaking style and manner.

Acoustic analysis indicates that stop-like /ð/ is significantly different from /d/ in burst amplitude, burst spectrum shape, burst peak frequency, and second formant at following-vowel onset. Moreover, the acoustic differences indicate that the dental place of articulation is preserved for stop-like /ð/. Automatic classification experiments involving these acoustic measures suggest that they are robust in distinguishing stop-like /ð/ from /d/. Applications of these findings may lie in areas of automatic speech recognition, speech transcription, and development of acoustic measures for speech disorder diagnosis.

Thesis Supervisor: Kenneth N. Stevens
Title: Clarence J. LeBel Professor of Electrical Engineering
and Professor of Health Sciences and Technology

Page intentionally left blank

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I. INTRODUCTION

The dental fricative /ð/, often found in word-initial positions of function words such as "*th*e" and "*th*is", is one of the most frequently spoken word-initial phonemes in English (Denes, 1963). However, it is also highly vulnerable to contextual modifications, especially in casual speech (Gimson, 1989; Jurafsky et al., 1998; Roach, 2000; Bell et al. 2003; Shockey, 2003). In word-initial position, studies have shown that this phoneme may take on characteristics of the preceding consonant; for example, it may become nasal when preceded by /n/, sonorant and lateral when preceded by /l/, or turn into a stop-like consonant when preceded by /t/ (Manuel, 1995; Manuel and Wyrick, 1999; Cao, 2002).

Stop-like /ð/ is a phonetic variation that deserves particular attention. As shown in **Figure 1.1**, the stop-like /ð/ exhibits a drastic modification from the canonical /ð/; its manner of production changes from one that is fricative to one that is stop-like. Most likely due to this change in production manner and the uncertainty over its place of articulation, stop-like /ð/ is sometimes transcribed as an alveolar stop [d][1] (Jurafsky et al., 1998). Thus, there is a need to determine whether the place of articulation is modified along with the manner of production for stop-like /ð/. In addition, the contextual frequency of stop-like /ð/ will provide further insights into the mechanisms that underlie the modification. Such knowledge would not only add to our understanding of speech variations, but could also be useful in applications such as speech transcription, automatic speech recognition, and the diagnosis of speech disorders.

[1] The brackets indicate that a segment is phonetically transcribed as [d], while the actual phoneme may be something else.

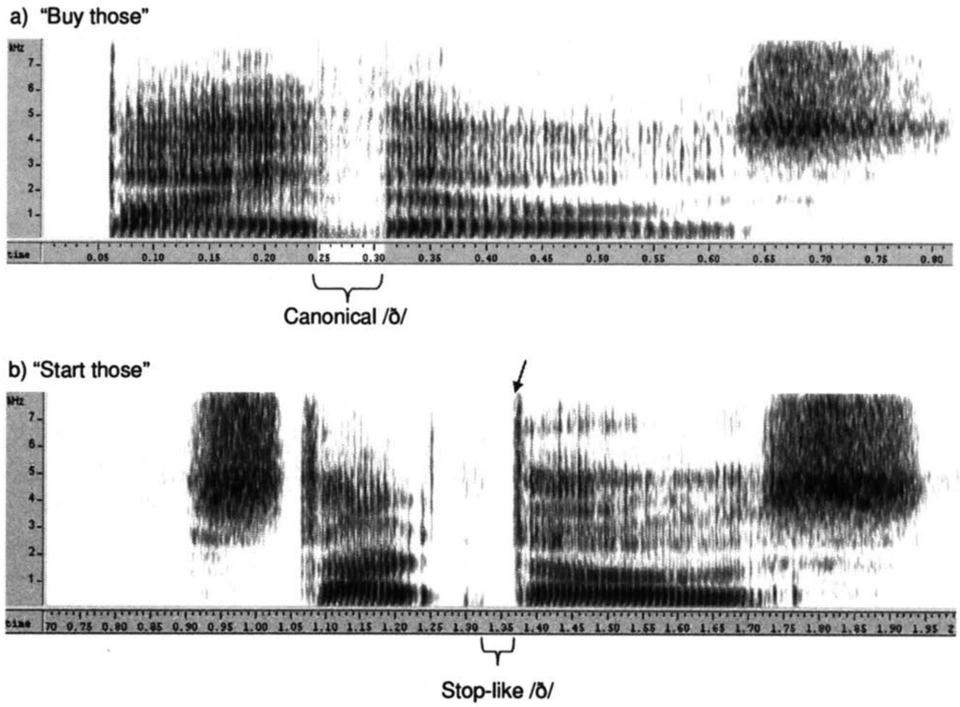


Figure 1.1: Spectrograms of a canonical /ð/ and a stop-like /ð/. Utterances are produced by a male speaker. a) Spectrogram of the canonical /ð/ in "buy those". b) Spectrogram of a stop-like /ð/ in "start those". The canonical /ð/ has voicing and weak noises in frequencies above 3 kHz throughout the entire segment. On the other hand, stop-like /ð/ contains a period of no acoustic energy, suggesting that a complete closure is formed in the oral cavity, followed by a sudden acoustic burst upon the release of the closure.

This thesis research involves three major studies. The first study aims to gain insight into the frequency of stop-like /ð/ and the segmental contexts where it occurs. The major questions this study attempts to answer are:

- In what types of segmental contexts does stop-like /ð/ occur? What is the frequency of the occurrence?
- Does the frequency of stop-like /ð/ differ for read speech, task-oriented spontaneous speech, and non-task-oriented spontaneous speech?

- What insights might the contextual frequency information provide as to the mechanisms that underlie the stop-like modification of /ð/?

The second study involves detailed acoustic analysis of stop-like /ð/. Comparisons are made with /d/, a consonant that is sometimes confused with stop-like /ð/. Major questions addressed by this study are:

- Does acoustic evidence suggest that the dental place of articulation is preserved for stop-like /ð/?
- If the dental place is preserved, then what are the salient acoustic cues for distinguishing stop-like /ð/ from /d/?

The third study consists of building an automatic classifier for stop-like /ð/ and /d/. The following questions are explored:

- Are the salient acoustic cues obtained in the second study of this thesis robust enough to be used for automatic classification of stop-like /ð/, distinguishing it from /d/?
- What are the applications of this research for automatic speech recognition systems?

Based on past research findings, it is hypothesized that stop-like /ð/ will occur most often when preceded by a stop consonant and notably less often in other contexts. The frequency of this stop-like modification is expected to increase with the level of casualness of the speech; more specifically, it is expected to be the lowest for read speech and highest for non-task-oriented spontaneous speech. The dental place of /ð/ is expected to be preserved despite the change in manner of production. Therefore, it is hypothesized that acoustic cues that distinguish between stop-like /ð/ and /d/ should exist and should provide a basis for building effective automatic classifiers of the consonants.

The next chapter will present the motivation and significance of this thesis research. It is followed by an overview of the acoustics and modifications of /ð/, as well as a review of relevant past studies in Chapter III. Then, the three major studies of this thesis will be presented and discussed in detail in Chapters IV, V, and VI. Applications of this research in automatic speech recognition and the general study of variations in speech, along with future directions in the study of stop-like /ð/, will be explored in Chapters VII, VIII, and IX before concluding remarks in Chapter X.

II. MOTIVATION AND SIGNIFICANCE

Phonetic variations occur widely in speech. Many of the most common variations are still shrouded in mystery, in terms of the nature and extent of their change from the canonical forms. Some of these variations pertain directly to the dental fricative /ð/, a phoneme that frequently occurs in spoken English, most often in word-initial positions of function words (Denes, 1963; Maddieson, 1984). In casual speech, /ð/ is frequently modified from its full form (Lee, 1990; Jurafsky et al., 1998; Bell et al. 2003). For example, /ð/ in the function word 'the' may be transcribed as [ð], [dh], [d], or [θ]. It may also become nasalized, lateralized, stop-like, or omitted all together (Cao, 2002; Glass, 1998; Lee, 1990; Manuel, 1995; Manuel and Wyrick, 1999).

The stop-like variant of /ð/, known to commonly occur when preceded by a stop consonant, is the focus of this thesis work. This variation is particularly interesting because the manner of production is drastically changed for /ð/, from one that is fricative to one that is stop-like. Listeners seem to have no problems in perceiving the stop-like variant of /ð/ as a /ð/. What characteristics of the stop-like form still make the consonant recognizable as a /ð/? Is the dental place of articulation preserved despite the change in manner of production? Uncertainly over these questions sometimes lead to the confusion of stop-like /ð/ with /d/. Therefore, these questions, along with ones regarding the contextual frequency of stop-like /ð/, need to be answered in order to better understand the nature of this common variation.

Such knowledge will be immediately useful for applications in speech transcription and automatic speech recognition. For example, information on how stop-like /ð/ may be acoustically distinguished from /d/ can be used to label /ð/ with better accuracy. Improving transcription accuracy may also benefit automatic speech recognition (ASR) systems, since most commercial systems are trained on labeled speech data (Young, 1996; Young 2001; Glass and Zue, 2005). Furthermore, feature-based ASR systems can benefit from knowing which acoustic cues to use in order to effectively extract the features of /ð/ despite modification. In addition, the features for this particular phoneme may be ranked according to their robustness in the face of modification. For example, if it is found that the dental place is always preserved for /ð/ despite modification, then features associated with place of articulation may be ranked as being more robust than those associated with the manner of production for that particular phoneme. Speech recognition systems that employ probabilistic information on when a sound is most likely to occur can use contextual frequency data on stop-like /ð/ to improve detection accuracy, especially when training data for the system is sparse.

Furthermore, the results of this research may be useful in the diagnosis and treatment of speech disorders. For example, if a patient always produces a stop-like /ð/ instead of a continuant one, then it is important to know which acoustic cues to examine in order to deduce the patient's place of articulation for the consonant. That way, the clinician can determine whether the patient has trouble with the place of articulation as well as the manner of production for /ð/. These pieces of information would be essential in the correct diagnosis and eventual treatment of a speech problem.

Finally, the research approach of this thesis, which centers on detailed acoustic analysis of the speech signal, may potentially be generalized and modified to study other types of speech variations. Since the speech signal is a manifestation of production mechanisms, the non-invasive method of detailed acoustic analysis is one effective way to uncover answers to when, why, and how certain phonetic modifications occur. It is hoped that in addition to gathering knowledge specific to /ð/, this research may give insights in the general study of variations in speech.

III. ACOUSTICS AND MODIFICATIONS OF /ð/

This section provides an overview of the production and acoustics of /ð/, followed by a discussion of the expected acoustic differences between a dental and an alveolar stop consonant. Past research findings on modifications of /ð/ are also reviewed.

Production and Acoustics of /ð/

/ð/ is a voiced, non-strident, dental fricative with weak noise. Its production involves vocal fold vibration as well as air turbulence through the vocal tract constriction formed by the tongue between the upper and lower teeth; or the tongue behind the back of the upper teeth. Since there is no complete closure in the vocal tract during its production, /ð/ is continuant; there is continuous acoustic output throughout its production. On the other hand, stop consonants such as /d/ and /g/ are not continuant. The production of stop consonants involves the formation of an oral-cavity closure, which leads to a signal that contains a period of no acoustic output from the mouth. This period is followed by a sudden burst of acoustic energy upon the release of the oral-cavity closure.

Figure 3.1 displays the vocal tract configuration, as well as the tube approximation for the dental fricative. The oral configuration for /ð/ includes a short front cavity, with a spectral peak location around 7 to 8 kHz (Jongman, Wayland, and Wong, 2000; see Appendix A for Tube Approximation of /ð/). As **Figure 3.1(b)** illustrates, the noise source is close to the constriction for /ð/; this location is less effective in exciting the front cavity resonance than a source located further from the constriction (Stevens, 1998;

see Appendix B for detailed explanation of the excitation of tubes). In addition, there is no obstacle in the front cavity of /ð/ to create further turbulence for the rapid airflow through the dental constriction (Shadle, 1985; Pastel, 1987). Thus, there is relatively low amplitude in high frequencies for /ð/.

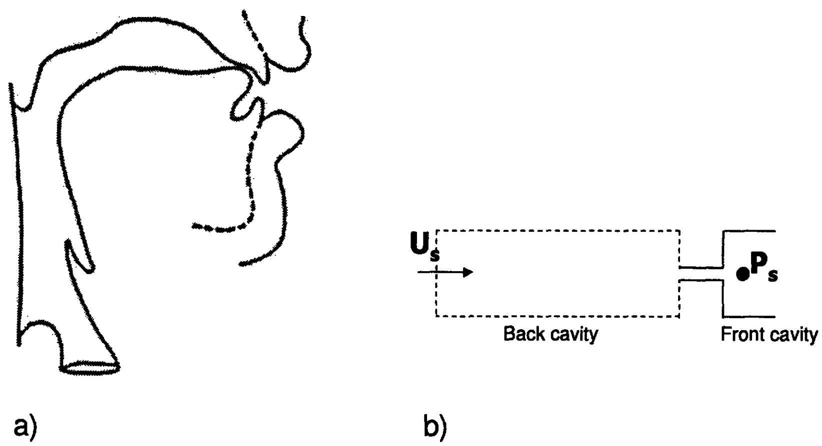


Figure 3.1: Configuration of a dental fricative. a) Mid-sagittal cross-sections of a vocal tract for a dental fricative (traced by Dennis Klatt from Perkell's (1969) film). b) Tube approximation of the dental fricative. U_s represents the glottal source and P_s represents the turbulent noise source. The dotted line of the back cavity indicates that it is much longer than what is scaled in the drawing.

Modifications of /ð/

As previously mentioned, /ð/ has been found to undergo a number of phonetic modifications. **Figure 3.2** contains spectrograms of a canonical /ð/, a nasalized /ð/, a sonorant /ð/, and a stop-like /ð/. The spectrograms are obtained from utterances of a male speaker, recorded at the Massachusetts Institute of Technology (MIT) Research Lab of Electronics Speech Communication Group. In all of these modified cases, the fricative manner of production for /ð/ is changed. However, past studies have found acoustic evidence suggesting that the dental place of articulation is preserved despite these modifications.

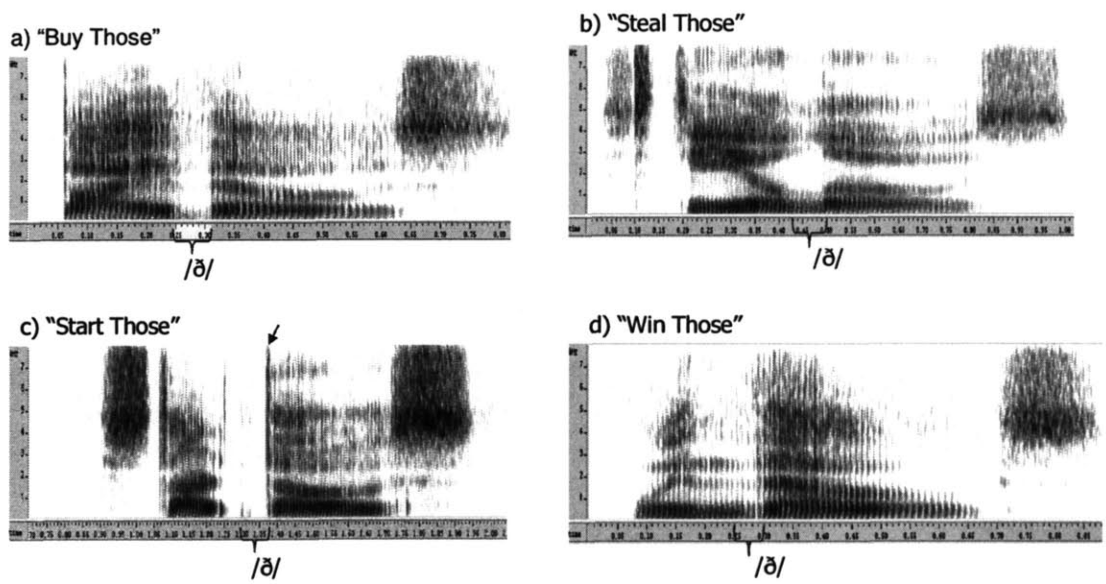


Figure 3.2: Spectrograms of canonical and modified /ð/. Utterances are produced by a male speaker. The canonical form in "buy those" is shown in a). Sonorant and lateralized /ð/ in "steal those" is shown in b). Stop-like /ð/ in "start those" is shown in c). Nasalized /ð/ in "win those" is shown in d).

Manuel's (1995) study of the spectral characterization of /ð/ preceded by /n/, in utterances produced by six speakers, reveals that although the dental fricative has become nasalized, acoustic evidence still indicates that the dental place of articulation is retained. That evidence includes lower F2 at vowel onset following the release of nasalized /ð/ than following the release of /n/, in the same context. The lower F2 is expected if the dental place is preserved for nasalized /ð/; this is due to the more backed tongue body and longer back cavity for a dental than for an alveolar place of articulation (Stevens and Keyser, 1989). Manuel's perception tests involving synthetic stimuli with varying F2 frequencies at the nasal consonant regions of 'win noes' and 'win those' have revealed that listeners report hearing 'those' more often when F2 is low at the nasal consonant release. This suggests that F2 could be used by listeners as a salient cue for distinguishing contextually-nasalized /ð/ from /n/.

Another study centers on cases of /ð/ which have taken on characteristics of the preceding /l/ (Manuel and Wyrick, 1999). From the acoustic analysis of /ð/ in utterances produced by eight speakers, evidence indicates that the dental fricative becomes sonorant and lateral when preceded by /l/. However, acoustic data indicate that F2 measured at the release of the modified /ð/ is higher than that at the release of /l/, but lower than that of a true /ð/. The higher F2 measured for /ð/ is expected, since the dental consonant would typically be made with a more fronted tongue body than for /l/. The data suggest that the dental place of articulation is likely retained for /ð/. Perception tests reveal that when the modified /ð/ is heard out of context, listeners almost always perceive 'Loew's' instead of 'those'. However, when heard in the context of 'steal those', listeners almost always perceive 'those'. These results suggest that context plays an important role in the perception of /ð/.

More recently, Cao (2002) has examined the acoustic characteristics of /ð/ when preceded by /n/, /t/, or an English vowel, in phrases spoken by four speakers. She observes that /ð/ is almost always produced as a stop-like consonant when the preceding consonant is a /t/. This finding agrees with previous observations where word-initial /ð/ is stop-like when preceded by a stop consonant (Roach, 2000; Shockey, 2003). In addition, Cao finds that for the two female speakers, stop-like /ð/ has lower high-to-midfrequency amplitude ratio in its stop burst and lower second formant (F2) at the onset of its succeeding vowel than for /d/, in the same segmental context; these findings suggest that the dental place of articulation is retained for stop-like /ð/. However, the differences are not apparent between stop-like /ð/ and /d/ tokens produced by the two male speakers in Cao's study. Therefore, it is unclear as to whether the preservation of the dental place can be generalized across speakers for stop-like /ð/.

Acoustic Differences between Dental and Alveolar Stops

As previously mentioned, stop-like /ð/ involves a change from a fricative manner of production to one that is stop-like. If the dental place of articulation is preserved for stop-like /ð/, then it should exhibit the acoustic characteristics of a dental stop consonant rather than an alveolar stop consonant.

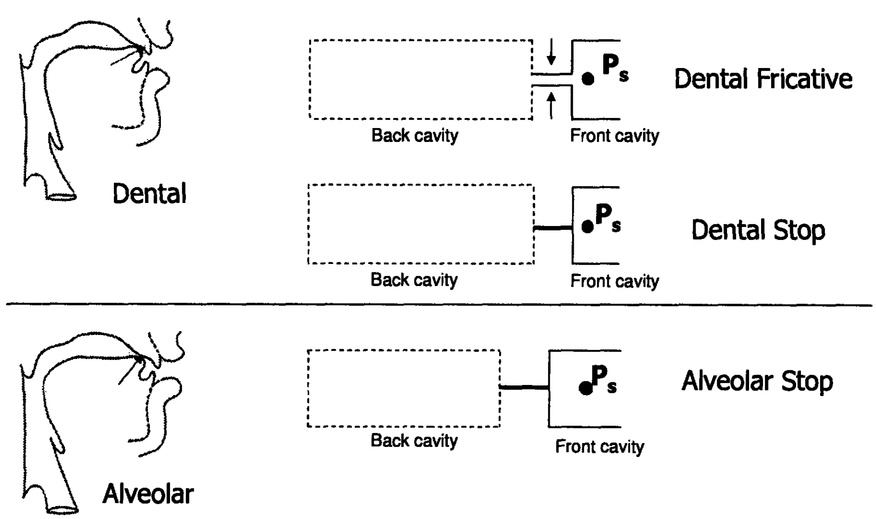


Figure 3.3: Tube approximations of a dental fricative, dental stop, and alveolar stop. Mid-sagittal cross-sections of a vocal tract for a dental obstruent and an alveolar obstruent are shown with their respective tube approximations. The constriction point highlights a main difference between the two places of articulation. Moreover, it is noted that a dental stop is produced with a complete closure while a dental fricative is produced with a non-zero constriction.

Figure 3.3 shows tube approximations for a dental stop and an alveolar stop consonant. The expected acoustic differences between the two types of obstruents are listed in **Table 3.1**. The closure for an alveolar stop is made with the tongue blade against the alveolar ridge, whereas that for the dental is made against the upper teeth. Therefore, the front cavity of a dental stop is shorter than that for an alveolar stop. Shorter tubes have higher resonant frequencies than longer tubes (Stevens, 1998; also see Appendix A for tube approximation). Thus, when the front cavity is excited by a noise source at the release of the complete closure, the burst amplitude peak for the dental should be higher in frequency than that for the alveolar.

Acoustic Cue	**Expected Difference**	**Reason**
Burst Spectrum Peak	Higher for Dental	Dentals have a shorter front cavity than alveolars
Burst Amplitude	Lower for Dental	Dentals have a noise source that is both weaker and in a worse location for exciting the shorter front cavity than for alveolars
Burst Spectrum Shape	Flatter for Dental	
F2 at Onset of Following Vowel	Lower for Dental	Dentals have a more backed tongue body and a longer back cavity

Table 3.1: The expected acoustic differences between a dental and an alveolar obstruent.

Furthermore, there is no obstacle in the front cavity to create further turbulence when rapid airflow exits the dental constriction. The alveolar configuration, on the other hand, has the back of the teeth as an obstacle in the front cavity. Thus, the amplitude of its noise source is increased when rapid airflow through the constriction directly impinges on the back of the teeth (Shadle, 1985; Pastel, 1987). The noise source is not only weaker for the dental stop, it is also located closer to the constriction, a position that is less effective in exciting the front cavity (Stevens, 1998; see Appendix B for detailed explanation of the excitation of tubes). Therefore, high frequency amplitude as a result of front-cavity excitation should be lower and flatter in shape for the dental than for the alveolar. Finally, the tongue body is expected to be lower and more backed for the dental than for the alveolar; this should lead to lower second formant (F2), a back-cavity resonant frequency, at the onset of a vowel following a dental consonant than that at the onset of a vowel following an alveolar consonant (Stevens and Keyser, 1989).

These theoretically-based acoustic differences between dental and alveolar obstruents are supported by findings of previous studies. Acoustic comparisons of dental and alveolar stop consonants have been carried out in studies of several different languages. In a study on the acoustic properties of dental and alveolar stop consonants in Malayalam, the root-mean-square (rms) amplitude of the stop consonant burst normalized to the rms of the adjacent vowel amplitude is found to be salient in distinguishing between the two places of articulation (Jongman, Blumstein, and Lahiri, 1985). The normalized amplitude of the alveolar stop burst is consistently higher than that of the dental stop burst. This acoustic attribute is used to distinguish dental stops from alveolar stops in Malayalam with 91.8% accuracy. However, it is not as successful in distinguishing between dental stops in Dutch and alveolar stops in American English; 68.2% of the American-English alveolar stops and 63.2% of the Dutch dental stops are classified correctly. Even though American-English alveolar stops are characterized by stronger bursts than Dutch dental stops, speaker-to-speaker variability within each of the two languages seems to have made accurate classification based on rms burst amplitude difficult. The authors suggested that the greater variability may be explained by the non-contrastive nature of a dental-alveolar distinction in stop consonants of Dutch or English.

Stoel-Gammon and colleagues (1994) have found that American-English alveolar /t/ and Swedish dental /t/ differed in voicing onset time (VOT), burst intensity and burst spectra. The VOT values of American-English /t/ are significantly longer than those of Swedish dental /t/. The burst intensity measures are made by taking the following vowel intensity maximum in dB and subtracting the stop-burst intensity maximum in dB. This measure is found to be lower for the American-English alveolar /t/ than for the Swedish dental /t/, with statistical significance. This indicates that the American-English alveolar /t/ has higher burst intensity than the Swedish /t/. In addition, the authors have studied the burst spectra for the two types of stops, using spectral moment analysis described by Forrest et al. (1988). The four spectral moments are used to obtain average energy concentration (mean frequency), spectral shape as indicated by the spread of frequency around the mean (standard deviation), tilt of the distribution (skewness), and degree of peakedness (kurtosis). The analysis indicates that, compared to Swedish dental /t/, American-English alveolar /t/ is significantly more compact and more peaked.

Sundara (2005) has examined the acoustic differences between Canadian English (CE) alveolar stops /d t/ and Canadian French (CF) dental stops /d t/. Voicing onset time (VOT) is found to be significantly longer for CE alveolar stops than for CF dental stops. Burst intensity is found to be stronger for CE bursts than for CF bursts. Burst spectral shape, measured in terms of mean frequency, standard deviation, and kurtosis, is also different between CE alveolar stops and CF dental dentals. Consistent with the findings of Stoel-Gammon et al., the alveolar stops have higher mean burst frequency, smaller standard deviation, and higher kurtosis than dental stops.

The acoustic differences found between dental and alveolar stops in these studies are consistent with the expected differences based on theoretical derivations. Therefore, if the dental place were preserved for stop-like /ð/, then it should have a higher burst spectrum peak frequency, lower burst amplitude, flatter burst amplitude shape, and lower F2 at the following vowel than the alveolar /d/.

After an examination of the segmental context of stop-like /ð/ in the next chapter, detailed acoustic analysis of the modified consonant, in comparison with /d/, will be presented in Chapter V. Emphasis is placed on whether the acoustic differences, if they exist, agree with what is expected if the dental place were preserved for stop-like /ð/. Attention is also paid to those acoustic cues that distinguish between stop-like /ð/ and /d/. The robustness of these acoustic cues in the automatic classification of stop-like /ð/ and /d/ will be discussed in Chapter VI. The remaining chapters will focus on the applications of this research in automatic speech recognition and the general study of variations in speech, along with future directions in the study of stop-like /ð/.

IV. CONTEXTUAL FREQUENCY OF STOP-LIKE /ð/

This study attempts to answer the following questions:

- In what types of segmental contexts does stop-like /ð/ occur? What is the frequency of its occurrence?
- Does the frequency of stop-like /ð/ differ for read speech, task-oriented spontaneous speech, and non-task-oriented spontaneous speech?
- What insights might the contextual frequency information provide as to the mechanisms that underlie the stop-like modification of /ð/?

Based on past research findings, it is hypothesized that stop-like /ð/ would occur most often when preceded by a stop consonant and notably less frequently in other contexts. In addition, the frequency of the stop-like modification is expected to increase with the casualness of the speech. More specifically, it is expected to be the lowest for read speech and highest for non-task-oriented spontaneous speech.

Research Methodology

The analysis of word-initial /ð/ from continuous speech is conducted for three different speech databases, the TIMIT database, the American-English Map Task (AEMT) Corpus, and the Buckeye Corpus.

The TIMIT database consists of utterances from 630 speakers of American English, covering all the major regions of the country (Fisher et al, 1987; TIMIT Documentation, 1990). The text corpus was designed by a joint effort among the Massachusetts Institute of Technology (MIT), Stanford Research Institute (SRI), and Texas Instruments (TI). The recordings were made at Texas Instruments in sound-attenuated chambers using headset microphones. For this study, utterances from speakers with no distinct regional accents, as indicated by the TIMIT documentation, are selected for analysis.

The AEMT Corpus contains 16 dialogues, each between two subjects as they work to complete a map-navigation task (MIT RLE Speech Communication Group AEMT Documentation). The task involves one person giving directions to the other in order to reach a destination on the map. There are a total of 8 female American-English speakers in the corpus; they grew up in various parts of the country. The recordings are made in a sound-attenuated chamber at MIT, using the Map-Task protocol (McAllister et al, 1990). All tokens of /ð/ from the 16 recordings obtained from the subjects giving directions are analyzed. The corpus contain phonetic and word transcriptions, as well as prosodic labels for 8 out of the 16 recordings.

The Buckeye Corpus contains conversational speech produced by 40 talkers from central Ohio (Pitt et al., 2005). Each talker is recorded through a head-mounted microphone while he or she is being interviewed by an experimenter. The recordings are made in a seminar room at Ohio State University. The subjects are told that they are participating in a study to learn how people express "everyday" opinions in conversations. It is indicated to them that the actual topic of their conversations is not important. The recording sessions consist of the interviewer eliciting opinions from the interviewee based on his/her background; it is reported that after 5 to 10 minutes of "eliciting" opinions from the subject, the interview would typically become a friendly conversation. For this study, recordings from 20 talkers, 11 female and 9 male, are analyzed. At the time of the thesis work, these are all of the transcribed and publicly available recordings.

The analysis begins with finding cases of /ð/ from the recordings of the various databases. This is done through a text search for "th" in the word labels of each selected recording. From that point on, only cases of word-initial /ð/ are examined acoustically. Word-initial /ð/ is examined because it is found in a variety of segmental contexts, whereas word-medial /ð/ and word-final /ð/ are always preceded by a vowel.

Criteria for Stop-like /ð/ Categorization

Word-initial /ð/ is categorized as stop-like when there is acoustic evidence to indicate that a complete closure is made and released. A stop-like /ð/ segment must contain one of the following series of acoustic events in the speech waveform:

- A period of silence followed by a release burst (see **Figure 4.1**);

  OR

- A decreasing voice bar followed by a period of silence before a release burst (see **Figure 4.2**);

  OR

- A decreasing voice bar followed by a release burst (see **Figure 4.3**).

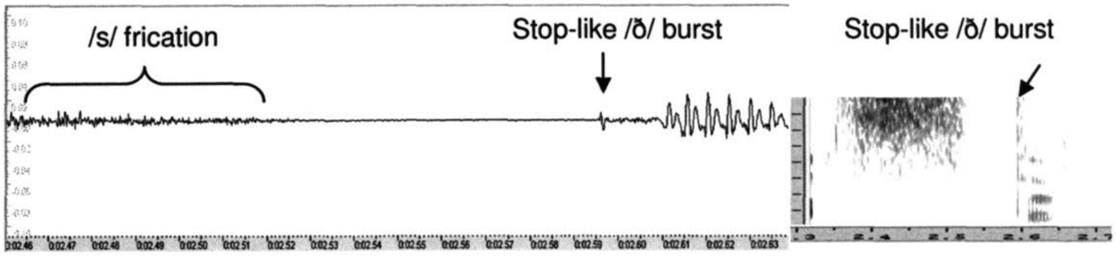


Figure 4.1: A stop-like /ð/ release that is preceded by a stop gap. /ð/ is preceded by /s/ in the utterance.

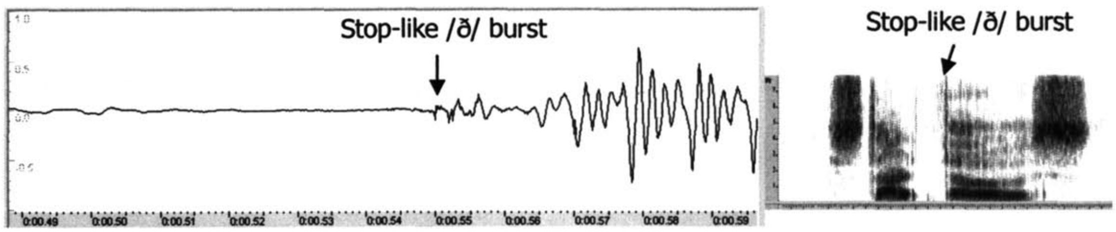


Figure 4.2: A stop-like /ð/ release that is preceded by a stop gap and some evidence for voicing. /ð/ is preceded by a /t/ with no release burst. There is some voicing prior to the stop gap.

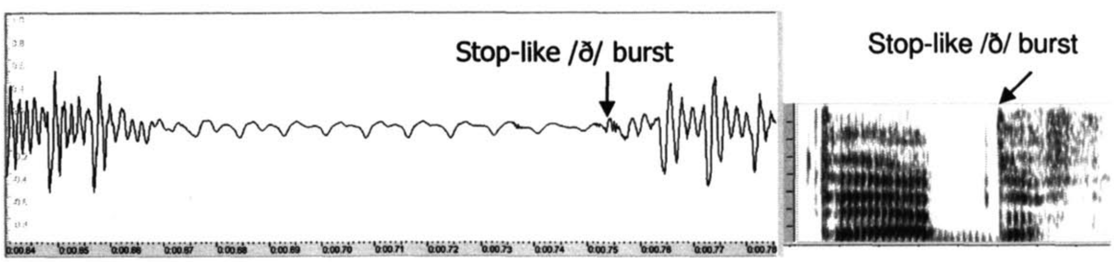


Figure 4.3: A stop-like /ð/ release that is preceded by a decreasing voice bar. /ð/ is preceded by /b/. There is no stop gap in this /ð/ segment.

Stop-consonant production involves a complete closure in the oral cavity. Pressure builds up behind the closure, which leads to a sudden onset of acoustic energy upon the release of the closure. Therefore, a release burst must always be present in order to characterize a /ð/ token as stop-like in production. In addition, the release burst must be preceded by a period of silence (see **Figure 4.1**) or a decreasing voice bar (see **Figure**

4.3). The period of silence that typically precedes a release burst is known as a stop gap; the lack of acoustic output from the mouth indicates that a complete closure is formed in the oral cavity. In some cases of stop consonants, a decreasing voice bar, instead of a stop gap, precedes the release burst. This occurs when there is vocal fold vibration during the period of oral-cavity closure. The complete oral closure will lead to the decrease in the pressure drop across the glottis, thereby decreasing the amplitude of the voice bar. There are cases where a decreasing voice bar precedes a stop gap, followed by a release burst (see **Figure 4.2**).

Figure 4.4 contains examples of non-stop-like /ð/, one in utterance-internal position and another in utterance-initial position. Additional examples of stop-like and non-stop-like cases of /ð/ are included in Appendix D.

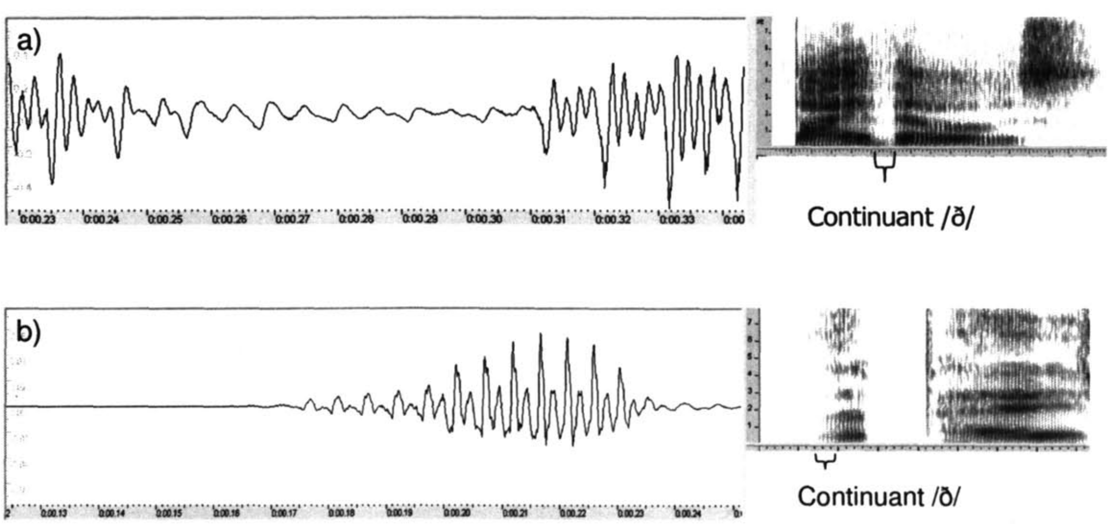


Figure 4.4: Examples of non-stop-like /ð/. Top panel (a) illustrates an utterance-internal /ð/ in "buy those" and bottom panel (b) an utterance-initial /ð/ in "the pa(tient)". Note that the two examples are not categorized as being stop-like; they do not satisfy the stop-like criteria and do not seem to contain evidence for complete closure.

For each database, /ð/ tokens are grouped according to segmental context. These contexts include the word-initial /ð/ preceded by silence, a consonant (across word boundary), or a vowel (across word boundary). The contexts are obtained from the phonetic labels used in each database and double checked by the author through listening to the speech recordings.

Contexts with preceding nasal consonants are not part of this study. For such contexts, it would be highly likely for /ð/ to become nasalized and have high energy at low frequencies (Manuel, 1995). Even if a complete oral closure is formed for a nasalized /ð/, there would be no pressure build-up in the oral cavity due to the opening of the velopharyngeal port. Therefore, a nasalized /ð/ that is produced with complete oral closure may be expected to exhibit continuous voicing and acoustic output from the nose. Furthermore, the release of a complete oral closure for a nasalized /ð/ may not lead to an acoustic burst because there is no intra-oral pressure build-up. The examination of such contexts, which require a different set of stop-like categorization criteria, will not be pursued in this thesis.

Results

Stop-like /ð/ in the TIMIT Database

Analysis is conducted on 492 cases of word-initial /ð/, extracted from 1000 utterances produced by 26 female speakers and 64 male speakers in the TIMIT database. The speakers are either from the western region of the United States (TIMIT Directory 7) or moved around the country during childhood (TIMIT Directory 8); therefore TIMIT categorizes those speakers as not having distinct regional accents.

The word-initial /ð/ cases examined in this study are all part of function words. **Table 4.1** lists the function words and the number of their occurrences in the utterances from TIMIT directories 7 and 8. Out of the 492 cases studied, 241 cases are in "the", followed by 127 in "that", and 39 in "they".

	Occurrence
The	241
That	127
They	39
This	24
Their	12
Them	10
Those	10
There	8
Then	8
Than	6
These	5
Though	1
Thus	1

Table 4.1: Function words containing word-initial /ð/ in TIMIT directories 7 and 8.

Table 4.2 lists all the contexts included in the analysis. The segmental contexts of /ð/ are grouped into broader categories because there is a big variation in the number of cases per context; this variation ranges from 2 to 113 tokens per segmental context. These broad categories of segmental contexts include word-initial /ð/ preceded by silence (silence#), by a stop consonant (stop#), by a fricative or an affricate (fricative#), by a vowel (vowel#), and by a liquid consonant (liquid#). Each category has a sample size of greater than 30 /ð/ tokens. Based on the central limit theorem, a sample size of at least 30 is widely used as a rule of thumb to obtain statistically meaningful results (Sanders, 1998; iSixSigma Staff, 2003).

Context of /ð/	**Number of Cases**
/ð/ preceded by silence	*184*
/ð/ preceded by a stop consonant	*175*
k#	113
t#	38
d#	17
p#	5
b#	2
/ð/ preceded by a fricative or an affricate	*67*
v#	26
z#	14
s#	13
ð#	5
f#	4
j#	3
ch#	2
/ð/ preceded by a vowel	*35*
schwa#	11
i#	9
u#	4
o#	4
ɑi#	3
ao#	3
ei#	2
/ð/ preceded by a liquid	*31*
r#	22
l#	9

Table 4.2: /ð/ tokens per context in TIMIT directories 7 and 8, grouped according to frequency.

Word-initial /ð/ preceded by silence is the most common context in the TIMIT recordings analyzed; most of them are utterance-initial cases. Among contexts with preceding stop consonants, /ð/ is most often preceded by /k/. This high occurrence is mainly due to the fact that there are two sentences that are read by every speaker in TIMIT; one of the two sentences includes a k#ð sequence in 'like that'. In addition to being preceded by /k/, /ð/ is also frequently preceded by /t/ and /d/ in TIMIT directories 7 and 8. Few cases of /ð/ are found to be preceded by other stop consonants.

Among contexts with preceding fricatives or affricates, /ð/ is most often preceded by /v/, /z/, and /s/, respectively. When /ð/ is preceded by a vowel, that vowel is most often a schwa or /i/. Among contexts with preceding liquids, /ð/ is more often preceded by a /r/ than a /l/.

Figure 4.5 illustrates the number of /ð/ cases that are stop-like or not stop-like, grouped according to five broad categories of segmental contexts. Of the 184 /ð/ cases that are preceded by silence, 142 (77%) are stop-like. Stop-like /ð/ occurrence is 73% (127 out of 175 cases) when the preceding consonant is a stop. Out of the 67 /ð/ cases with a preceding fricative or affricate, stop-like occurrence is 28%. For the 35 cases preceded by vowels, 4 (11%) are stop-like. Out of 31 cases of /ð/ with a preceding liquid consonant, 5 (16%) are categorized as stop-like. The distributions of stop-like /ð/ grouped according to specific segmental contexts are listed in **Table 4.3**.

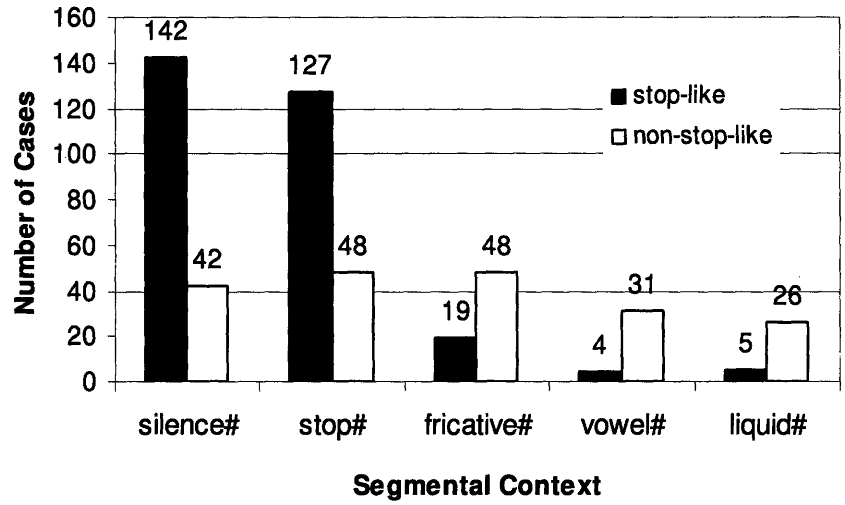


Figure 4.5: Stop-like /ð/ occurrence in TIMIT directories 7 and 8, grouped by categories of segmental context. These categories include /ð/ preceded by silence (silence#), by a stop consonant (stop#), by a fricative or an affricate (fricative#), by a vowel (vowel#), and by a liquid consonant (liquid#).

This analysis has also found that t#ð and d#ð sequences frequently consist of a stop-like /ð/ with no release of the preceding alveolar stop. Out of the 32 stop-like /ð/ cases preceded by /t/, 28 contain no /t/ release. For the 10 stop-like cases that are preceded by /d/, 7 contain no /d/ release. The lack of /k/ release burst is observed in 29 out of 82 k#ð sequences where /ð/ is stop-like. The absence of the preceding stop consonant's release burst suggests that the articulators may have formed the stop-like-/ð/ closure before the release of the preceding stop consonant, thereby obliterating the release burst of the preceding stop consonant. Thus, there seems to be articulatory overlap when stop-like /ð/ is produced in contexts where the preceding phoneme is a stop consonant.

Context of /ð/	Total Number of /ð/ Tokens	Stop-like /ð/ Tokens
/ð/ preceded by silence	*184*	*142(77%)*
/ð/ preceded by a stop consonant	*175*	*127 (73%)*
k#	113	82 (73%)
t#	38	32 (84%)
d#	17	10 (59%)
p#	5	2 (40%)
b#	2	1 (50%)
/ð/ preceded by a fricative or an affricate	*67*	*19 (28%)*
v#	26	6 (23%)
z#	14	3 (21%)
s#	13	4 (31%)
ð#	5	4 (80%)
f#	4	1 (25%)
j#	3	1 (33%)
ch#	2	0
/ð/ preceded by a vowel	*35*	*4 (11%)*
schwa#	11	1 (9%)
i#	9	0
u#	4	0
o#	4	0
ɑi#	3	2 (66%)
ao#	3	1 (33%)
ei#	2	0
/ð/ preceded by a liquid	*31*	*5 (16%)*
r#	22	3 (14%)
l#	9	2 (22%)

Table 4.3: Stop-like /ð/ tokens per context in TIMIT directories 7 and 8. For each context, the total number of /ð/ is listed with the number of stop-like cases.

The possible relationship between the stop-like modification of /ð/ and the voicing of its preceding phoneme is also investigated. Stop-like /ð/ occurrence is 69% (121 out of 175 cases) when the preceding obstruent consonant is voiceless, compared to 37% (25 out of 67 cases) when the preceding obstruent consonant is voiced. In addition, stop-like /ð/ rarely occurs (14% or 9 out of 66 cases) when the preceding phoneme is a vowel or a liquid consonant.

Among contexts with preceding stop consonants, stop-like /ð/ occurrence is 74% (116 out of 156 cases) when the preceding stop is voiceless and 58% (11 out of 19 cases) when the preceding stop is voiced. Correlation coefficient is calculated between the stop-like modification of /ð/ and the voicelessness of the preceding stop consonant (see Appendix E for the calculation of correlation coefficients). The resulting correlation coefficient is 0.35 from normalized data, indicating a weak positive correlation between the two variables. The calculation only included data from contexts with preceding stop consonants in order to limit other variables, such as the manner of production, which may also have effects on the relationship between the preceding phoneme and the stop-like modification of /ð/. It is noted that the normalized data for contexts with preceding /b/ and /p/ come from fewer than 10 cases of /ð/ each. Therefore, the small size of the data (fewer than the statistical rule of thumb of 30 samples) may not yield results that are as robust as ones from greater sample pools where a normal distribution may be assumed.

There does not seem to be a relationship between the stop-like modification of /ð/ and the voicing of the preceding fricative or affricate. Among contexts with preceding fricatives or affricates, stop-like /ð/ occurrence is 26% (5 out of 19 cases) when the preceding consonant is voiceless and 29% (14 out of 48 cases) when the preceding consonant is voiced. This finding is not consistent with the positive correlation between the stop-like modification of /ð/ and the voicing of a preceding stop consonant. It seems that factors other than or in addition to voicing may be involved.

In addition to the voicing of the preceding consonants, the place of articulation might also be correlated with the modification of /ð/. The data suggest that stop-like /ð/ frequently occurs when the preceding consonant is a coronal obstruent. For example, the stop-like modification occurs 84% (32 out of 38 cases), 80% (4 out of 5 cases), and 59% (10 out of 17) of the time when the preceding consonant is /t/, /ð/ and /d/ respectively. However, the stop-like modification of /ð/ and the preceding obstruent's coronal place of articulation seem weakly correlated. The correlation coefficient between the stop-like modification of /ð/ and the coronal place of articulation for the preceding consonant is 0.28, which indicates a weak positive correlation. The data indicate that the high frequency of stop-like /ð/ is not just limited to contexts with preceding coronal obstruents; the modification also occurs frequently when preceded by non-coronal obstruents.

Summary

The results from the TIMIT data reveal the following:

- High stop-like /ð/ occurrence is not exclusive to contexts where the consonant is preceded by a stop consonant.

- There is high occurrence (77% or 142 out of 184 cases) of stop-like /ð/ when preceded by silence; they are mostly in utterance-initial position.
- There is high occurrence of (73% or 127 out of 175 cases) of stop-like word-initial /ð/ when the preceding phoneme is a stop consonant.
- The stop-like modification occurs less frequently (28% or 19 out of 67 cases) when the preceding consonant is a fricative or affricate.
- This modification rarely (14% or 9 out of 66 cases) occurs when the preceding phoneme is a vowel or a liquid consonant.
- There is a weak positive correlation between the voicelessness of the preceding stop consonant and the stop-like modification of /ð/; however, there seems to be no correlation between the stop-like modification of /ð/ and the voicing of the preceding fricative consonant.
- There is weak correlation between the stop-like modification of /ð/ and whether the preceding consonant has a coronal place of articulation.
- The release burst of /t/, /d/, and /k/ are frequently absent when those consonants precede a stop-like /ð/. This suggests that the articulators may have formed a stop-like-/ð/ closure before the release of the preceding stop consonant.

Stop-like /ð/ in the American English Map Task Corpus

A total of 1561 cases of word-initial /ð/, from the 16 recordings of subjects giving directions in the AEMT Corpus, are analyzed. On average, the speakers uttered /ð/ every 5 seconds as they performed the map-navigation task.

The word-initial /ð/ cases examined in this study all occur in function words. **Table 4.4** lists the function words and the number of their occurrences in the 16 recordings. Out of the 1561 cases studied, 1032 occur in "the", followed by 176 in "then", and 175 in "that".

	Occurrence
The	1032
Then	176
That	175
There	59
This	17
Though	9
They	8
Those	7
Them	5
These	2
Than	1

Table 4.4: Function words containing word-initial /ð/ in the AEMT Corpus.

Table 4.5 lists all the segmental contexts of /ð/ in AEMT. Similar to the TIMIT database, the AEMT Corpus also has a big variation in the number of /ð/ per context; the sample size ranges from 1 to 235 tokens per context. This prompts the grouping of individual segmental contexts into the five broad categories of contexts mentioned in the previous section. The categories are word-initial /ð/ preceded by silence (silence#), by a stop consonant (stop#), by a fricative or an affricate consonant (fricative#), by a vowel (vowel#), and by a liquid consonant (liquid#). Each category has a sample size of greater than 80 /ð/ tokens.

Context of /ð/	Number of Cases
/ð/ preceded by a vowel	*406*
schwa#	235
o#	108
u#	25
i#	16
ɑ#	14
ɑi#	4
ao#	4
/ð/ preceded by a fricative or an affricate	*329*
v#	223
z#	52
s#	24
θ#	14
f#	9
ð#	5
sh#	1
j#	1
/ð/ preceded by a stop consonant	*152*
t#	100
k#	22
d#	19
p#	11
/ð/ preceded by silence	*110*
/ð/ preceded by a liquid	*82*
r#	73
l#	9

Table 4.5: /ð/ tokens per context in AEMT, grouped according to frequency. There are no cases where /b/, /g/, /zh/, or /ch/ precedes /ð/.Not included in the table are the 414 cases of /ð/ preceded by /n/, 51 cases preceded by /m/, and 28 cases preceded by /ng/; those cases are not acoustically analyzed in this study.

As shown in the table, /ð/ is most-frequently preceded by a vowel. Frequently occurring segmental contexts include a schwa preceding /ð/ in the phrase "to the" and an /o/

preceding /ð/ in phrases such as "below the" and "so the". Fricatives, affricates, and stop consonants are also found to frequently precede /ð/. For example, a frequent segmental context is /v/ preceding /ð/ in "of the" and "of that". Phrases such as "at the" and "that the", where /t/ precedes /ð/, often occur. Other common phrases include "is the", "pass the", and "over the". In addition, /ð/ is often preceded by a pause or breath, usually at the beginning of an utterance or smaller phrase.

It is noted that nasal consonants often precede /ð/. In fact, there are 414 cases of /ð/ that are preceded by /n/, spoken in phrases such as "in the", "on the", and "and then". Other commonly uttered phrases where a nasal precedes /ð/ include "from the" and "passing the". These tokens are not acoustically analyzed in this study.

Figure 4.6 illustrates the number of /ð/ tokens that are categorized as stop-like or not stop-like, grouped according to five broad categories of preceding-segmental contexts. Of the 110 /ð/ tokens that are preceded by silence, 87 (79%) are stop-like. Stop-like /ð/ occurrence is 88% (133 out of 152 tokens) when the preceding consonant is a stop. Out of the 329 /ð/ tokens with a preceding fricative or affricate, stop-like occurrence is 33% (109 tokens). For the 406 tokens preceded by vowels, 72 (18%) are stop-like. Out of 82 cases of /ð/ with a preceding liquid consonant, 13 (16%) are categorized as stop-like. The distributions of stop-like /ð/ grouped according to specific segmental contexts are listed in **Table 4.6**.

This analysis has also found that the /t/ release is often absent (85 out of 92 tokens) when preceding stop-like /ð/. The /d/ release is sometimes absent (7 out of 12 tokens) when it precedes a stop-like /ð/. The /k/ release is also sometimes absent (10 out of 20 tokens) when preceding stop-like /ð/. This, along with similar observations from the TIMIT recordings, suggests that the articulators may have formed a stop-like /ð/ closure before the release of the preceding stop consonant, thereby obliterating the release burst of the preceding stop consonant. Thus, there seems to be articulatory overlap when stop-like /ð/ is produced in contexts where the preceding phoneme is a stop consonant.

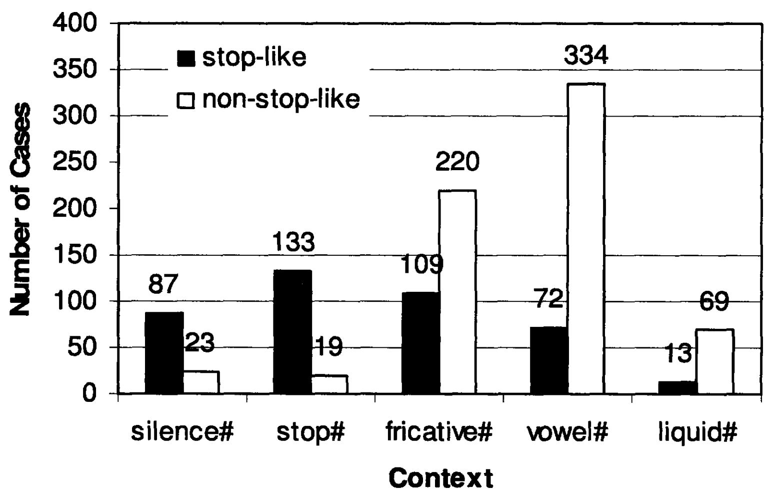


Figure 4.6: Stop-like /ð/ occurrence in AEMT, grouped by categories of segmental context. These categories include /ð/ preceded by silence (silence#), by a stop consonant (stop#), by a fricative or an affricate (fricative#), by a vowel (vowel#), and by a liquid consonant (liquid#).

A possible correlation between the voicing of the preceding obstruent and the stop-modification is investigated. Out of the contexts with a preceding stop, the lowest occurrence of stop-like /ð/ is when it is preceded by /d/, the only voiced stop in the set. For contexts with a preceding fricative or affricate, 85 out of 281 /ð/ cases (30%) are stop-like when the preceding consonant is voiced; whereas 24 out of 48 cases (50%) are stop-like when the preceding consonant is unvoiced. The correlation factor between the stop-like modification of /ð/ and the voicelessness of the preceding fricative or affricate is 0.1, indicating a weak positive correlation.

Context of /ð/	**Total Number of /ð/ Tokens**	**Stop-like /ð/ Tokens**
/ð/ preceded by a vowel	*406*	*72 (18%)*
schwa#	235	37 (16%)
o#	108	23 (21%)
u#	25	2 (8%)
i#	16	0
ɑ#	14	7 (50%)
ɑi#	4	1 (25%)
ao#	4	2 (50%)
/ð/ preceded by a fricative or an affricate	*329*	*109 (33%)*
v#	223	67 (30%)
z#	52	13 (25%)
s#	24	7 (29%)
θ#	14	12 (86%)
f#	9	4 (44%)
ð#	5	4 (80%)
sh#	1	1
j#	1	1
/ð/ preceded by a stop consonant	*152*	*133 (88%)*
t#	100	92 (92%)
k#	22	20 (91%)
d#	19	12 (63%)
p#	11	9 (82%)
/ð/ preceded by silence	*110*	*87 (79%)*
/ð/ preceded by a liquid	*82*	*13 (16%)*
r#	73	12 (16%)
l#	9	1 (11%)

Table 4.6: Stop-like /ð/ tokens per context in AEMT. For each context, the total number of /ð/ is listed with the number of stop-like cases.

The possible correlation between the stop-like modification of /ð/ and the preceding consonant's place of articulation is also explored. The data indicate that stop-like /ð/ occurs more often when the preceding consonant is coronal than when it is not. Out of the 214 tokens of /ð/ preceded by a coronal obstruent, 140 (65%) are stop-like. On the other hand, only 102 out of 267 tokens (38%) of /ð/ are stop-like when preceded by a non-coronal obstruent. The correlation coefficient between the stop-like modification of /ð/ and the coronal place of articulation of the preceding obstruent is 0.01; this suggests virtually no correlation between the two variables.

In addition to the segmental context, the prosodic context of /ð/ is also analyzed for 8 out of 16 AMET recordings that are prosodically labeled. **Figure 4.7** contains the number of stop-like /ð/ cases grouped according to prosodic context. Cases labeled as being merged with the preceding word (2-words merge) are the lowest on the prosodic hierarchy; higher up is the normal word boundary (Shattuck-Hufnagel and Turk, 1996). The "mixed case" includes those that seem higher up in the prosodic hierarchy than a normal word boundary, but lower than a small phrase boundary. The larger phrase boundary is the highest in the hierarchy among the prosodic labels. Out of the 49 cases of word-initial /ð/ labeled as "2-words merge", only 3 (6%) are stop-like. The highest number of word-initial /ð/ occurs at a normal word boundary, where 80 out of 348 (23%) are stop-like. There are 15 cases of /ð/ labeled as "mixed cases", 6 of which (40%) are stop-like. At the small phrase boundary, 6 out of 19 (32%) cases are stop-like. The highest percentage of stop-like /ð/ cases occur at the large phrase boundary, with 25 out 31 (81%) categorized as stop-like.

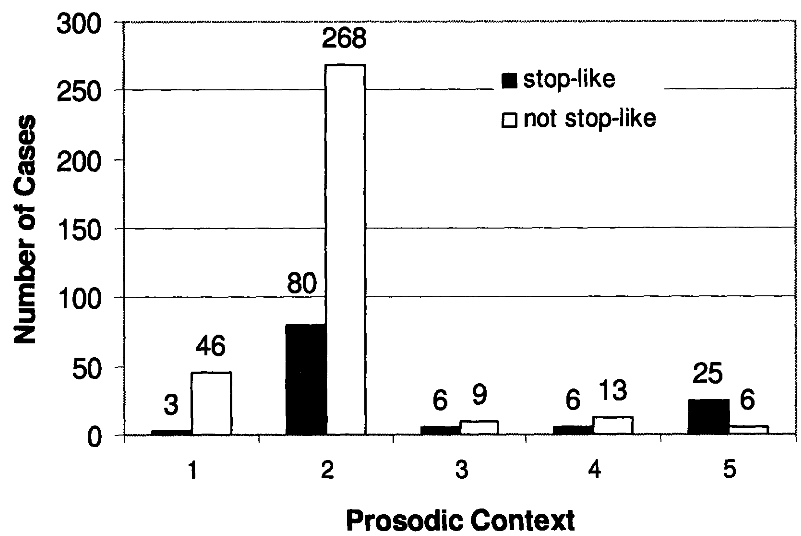


Figure 4.7: Prosodic context of /ð/ from AEMT

These occurrence rates are plotted in **Figure 4.8**. A generally upward trend in stop-like /ð/ occurrence is observed as the prosodic labels get higher on the prosodic hierarchy. There is a slight dip in the graph from "mixed case" to "small phrase boundary"; this may be due to the small number of total cases, 15 and 19, respectively, for each of the two contexts. In addition, it is noted that due to those small numbers, not enough cases are available to make the comparison across prosodic contexts while the segmental context remains the same.

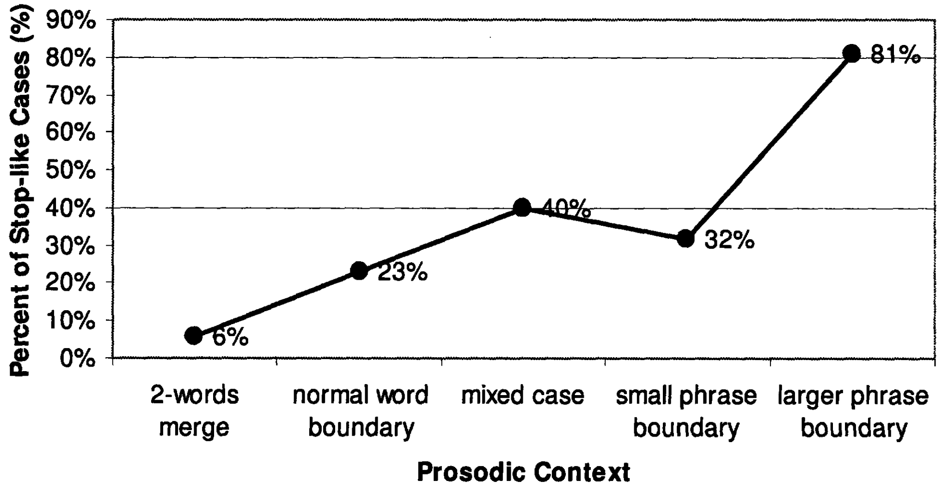


Figure 4.8: Percent of stop-like /ð/ per prosodic context.

Summary

The results from the AEMT data reveal the following:

- High stop-like /ð/ occurrence is not exclusive to contexts where the consonant is preceded by a stop consonant.
- When preceded by silence, /ð/ is frequently (79%) stop-like.
- Stop-like /ð/ frequently (88%) occurs when preceded by a stop consonant.
- The modification is less frequent (33%) when /ð/ is preceded by fricatives and affricates.
- The consonant is rarely (18%) stop-like when preceded by a vowel.
- Similarly, /ð/ is rarely (16%) stop-like when preceded by a /r/ or /l/.
- There may be a relationship between the voicelessness of the preceding stop consonant and the stop-like modification of /ð/; however, this relationship is not

clear due to d#ð being the only sequence in this corpus in which word-initial /ð/ is preceded by a voiced stop consonant.

- There is virtually no correlation between voicing of the preceding fricative consonant and the stop-like /ð/ modification.
- There is virtually no correlation between the stop-like modification of /ð/ and the preceding obstruent having a coronal place of articulation.
- The release burst of /t/, /d/, and /k/ are frequently absent when those consonants precede a stop-like /ð/. This suggests that the articulators may have formed a stop-like-/ð/ closure before the release of the preceding stop consonant.

These findings are consistent with those found for the read speech in the TIMIT database.

Stop-like /ð/ in the Buckeye Corpus

For this study, recordings from 20 talkers, 11 female and 9 male, are analyzed. For each talker, 48 to 50 sequentially-occurring /ð/ cases are analyzed for a randomly selected period of the conversation. A total of 998 cases of word-initial /ð/ are analyzed. Similar to the AEMT subjects, the talkers produce /ð/ at an average of every 5 seconds of the speech examined. The word-initial /ð/ cases examined in this study all occur in function words. **Table 4.7** lists the function words and their numbers of occurrence. Out of the 998 cases studied, 310 occur in "the", 275 occur in "that", and 151 occur in "they".

Table 4.8 lists the segmental contexts analyzed for /ð/ in Buckeye. Similar to TIMIT and AEMT databases, the Buckeye Corpus also has a big variation in the number of /ð/

tokens per segmental context. The sample size for each segmental context ranges from 2 tokens to 152 tokens. Therefore, individual segmental contexts are grouped into five broad categories of segmental contexts. They are /ð/ preceded by silence (silence#), by a stop consonant (stop#), by a fricative or an affricate (fricative#), by a vowel (vowel#), and by a liquid consonant (liquid#). Each category has a sample size of greater than 120 /ð/ tokens.

	Occurrence
The	310
That	275
They	151
There	93
This	49
Then	27
Them	26
Those	20
Their	19
Than	13
These	9
Therefore	3
Though	2
Themselves	1

Table 4.7: Function words containing word-initial /ð/ from Buckeye Corpus.

As shown in the table, /ð/ is most frequently preceded by a stop consonant in the Buckeye Corpus. A frequent segmental context is /t/ preceding /ð/, in phrases such as "about the" and "what the". Fricatives such as /v/ and /z/ are also found to frequently precede /ð/; they occur in phrases such as "of the" and "is the". A number of /ð/ tokens are found to follow vowels or liquids, in phrases such as "do the", "be the", and "for the". Furthermore, /ð/ is often preceded by a pause or breath, usually at the beginning of an utterance or smaller phrase. There are no cases where /b/, /g/, or /zh/ precedes /ð/.

Context of /ð/	Number of Cases
/ð/ preceded by a stop consonant	*269*
t#	152
k#	75
d#	39
p#	3
/ð/ preceded by a fricative or an affricate	*186*
v#	65
z#	64
s#	19
f#	15
ð#	14
θ#	3
ch#	3
j#	2
sh#	1
/ð/ preceded by a vowel	*152*
o#	31
i#	29
u#	26
schwa#	20
ɑ#	16
ao#	16
ʌ#	12
ɑi#	2
/ð/ preceded by silence	*139*
/ð/ preceded by a liquid	*120*
r#	60
l#	50

Table 4.8: /ð/ tokens per context in Buckeye, grouped according to frequency. Not included in the table are the 199 cases of /ð/ preceded by /n/, 26 cases preceded by /ng/, and 11 cases preceded by /m/; those cases are not acoustically analyzed in this study.

Nasal consonants are found to often precede /ð/. There are 199 cases of /ð/ that are preceded by /n/; they occur in phrases such as "in the", "and the", and "and then". There are also 26 /ð/ tokens preceded by /ng/ and 11 tokens preceded by /m/. Those cases of /ð/ are not acoustically analyzed in this study.

Figure 4.9 illustrates the number of /ð/ cases that are stop-like or not stop-like, grouped according to the five categories of segmental contexts. Of the 139 /ð/ tokens that are preceded by silence, 62 (45%) are stop-like. Stop-like /ð/ occurrence is 33% (89 out of 269 tokens) when the preceding consonant is a stop. Out of the 186 /ð/ tokens with a preceding fricative or affricate, stop-like occurrence is 9% (16 tokens). For the 152 tokens preceded by vowels, 9 (6%) are stop-like. Out of 120 cases of /ð/ with a preceding liquid consonant, 10 (8%) are categorized as stop-like. The distributions of stop-like /ð/ grouped according to specific segmental contexts are listed in **Table 4.9**.

This analysis also found that the /t/ release is frequently absent (55 out of 58 tokens), /d/ release is frequently absent (7 out of 9 tokens), and /k/ sometimes absent (12 out of 22 tokens), when preceding a stop-like /ð/. This, along with similar observations from the TIMIT and AEMT recordings, suggests that the articulators may have formed the stop-like-/ð/ closure before the release of the preceding stop consonant, thereby obliterating the release burst of the preceding stop consonant. Thus, there seems to be articulatory overlap when stop-like /ð/ is produced in contexts where the preceding phoneme is a stop consonant.

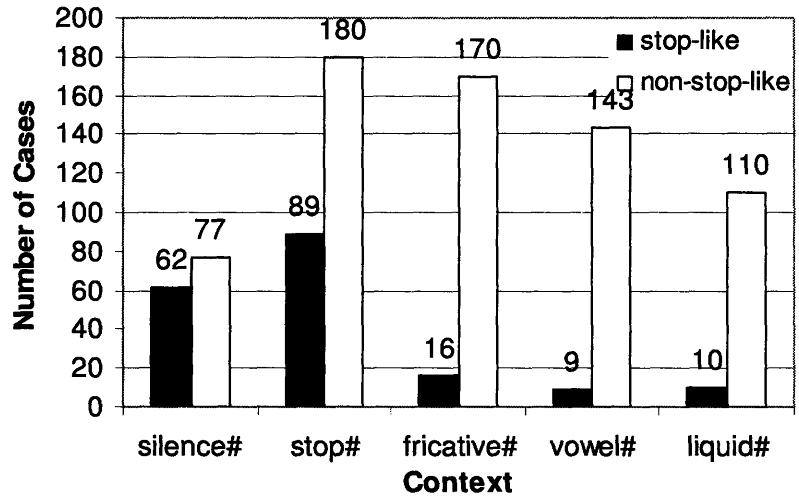


Figure 4.9: Stop-like /ð/ occurrence in Buckeye, grouped by categories of segmental context. These categories include /ð/ preceded by a stop consonant (stop#), by a fricative or an affricate (fricative#), by a vowel (vowel#), by silence (silence#), and by a liquid consonant (liquid#).

The possible correlation between the voicing of the preceding consonant and the stop-like modification of /ð/ is also investigated. It is found that 6 out of 41 cases (15%) of /ð/ are stop-like when preceded by a voiceless fricative or affricate, while 10 out of 145 (7%) are stop-like when preceded by a voiced one. As for stop consonants preceding /ð/, 80 out of 227 cases (35%) of /ð/ are stop-like when preceded by a voiceless stop, while 9 out of the 39 cases (23%) are stop-like when preceded by the voiced stop /d/. The correlation coefficient between voicelessness and stop-like occurrence is -0.01 for contexts with preceding stop consonants, indicating virtually no correlation. The correlation coefficient is 0.25 between the voicelessness of the preceding fricative or affricate and stop-like modification of /ð/, indicating a weak positive relationship.

Context of /ð/	Total Number of /ð/ Tokens	Stop-like /ð/ Tokens
/ð/ preceded by a stop consonant	*269*	*89 (33%)*
t#	152	58 (38%)
k#	75	22 (29%)
d#	39	9 (23%)
p#	3	0
/ð/ preceded by a fricative or an affricate	*186*	*16 (9%)*
v#	65	7 (11%)
z#	64	1 (2%)
s#	19	0
f#	15	5 (33%)
ð#	14	2 (14%)
θ#	3	1 (33%)
ch#	3	0
j#	2	0
sh#	1	0
/ð/ preceded by a vowel	*152*	*9 (6%)*
o#	31	1 (3%)
i#	29	2 (7%)
u#	26	1 (4%)
schwa#	20	3 (15%)
ɑ#	16	1 (6%)
ao#	16	1 (6%)
ʌ#	12	0
ɑi#	2	0
/ð/ preceded by silence	*139*	*62 (45%)*
/ð/ preceded by a liquid	*120*	*10 (8%)*
r#	60	7 (12%)
l#	50	3 (6%)

Table 4.9: Stop-like /ð/ tokens per context in Buckeye. For each context, the total number of /ð/ is listed with the number of stop-like cases.

In addition, the correlation between the stop-like modification of /ð/ and the preceding consonant's place of articulation is examined. The data indicate that stop-like /ð/ occurs slightly more often when the preceding consonant is coronal than when it is not. Out of the 291 tokens of /ð/ preceded by a coronal obstruent, 71 (24%) are stop-like. On the other hand, only 31 out of 164 tokens (19%) of /ð/ are stop-like when preceded by a non-coronal obstruent. The correlation coefficient between the stop-like modification of /ð/ and the coronal place of articulation of the preceding obstruent is smaller than 0.01; this suggests virtually no correlation between the two variables.

Summary

The results from the Buckeye Corpus data reveal the following:

- When preceded by silence, /ð/ is sometimes (45%) stop-like.
- Stop-like /ð/ sometimes occurs (33%) when it is preceded by stop consonants.
- The modification rarely occurs (less than 10%) when /ð/ is preceded by fricatives, affricates, vowels or liquids.
- There is virtually no correlation between voicing of the preceding stop consonant and stop-like modification of /ð/; a weak correlation is suggested for contexts with a preceding fricative or affricate.
- There is virtually no correlation between the stop-like modification of /ð/ and whether the preceding consonant has a coronal place of articulation.
- The release burst of /t/, /d/, and /k/ are frequently absent when those consonants precede a stop-like /ð/. This suggests that the articulators may have formed a stop-like-/ð/ closure before the release of the preceding stop consonant.

Overall, the occurrence of stop-like /ð/ is much lower in the non-task-oriented spontaneous speech of Buckeye than in the task-oriented spontaneous speech of AEMT or read speech of TIMIT. The highest occurrences of the modification happen when /ð/ is preceded by silence or a stop consonant; this is consistent with the results obtained from the TIMIT and AEMT. However, there is virtually no difference in stop-like /ð/ occurrence rates for contexts with preceding fricatives, affricates, vowels or liquids in the Buckeye recordings analyzed in this study.

Discussion

Differences and Similarities of /ð/ Behavior among the Databases

Stop-like /ð/ is found to occur, to varying degrees, in all types of contexts across the three databases. This finding does not support the hypothesis that the high occurrence of the stop-like /ð/ is exclusive to segmental contexts where the consonant is preceded by a stop consonant. **Table 4.10** lists the occurrence rates for stop-like /ð/ across the databases, grouped according to segmental context. Data from all three databases indicate that stop-like /ð/ occurs most frequently when preceded by silence or when preceded by a stop-consonant. The occurrence is much lower for contexts where the preceding consonant is a fricative or an affricate. The lowest occurrence rates for stop-like /ð/ are found for contexts where the preceding phoneme is a vowel or liquid consonant.

The findings of this study do not support the hypothesis that stop-like /ð/ would increase in frequency with the casualness of the speech, occurring most often in non-task-oriented speech in Buckeye and least often in read speech of TIMIT. Stop-like /ð/ is found to occur much more frequently in TIMIT and AEMT than in Buckeye.

	TIMIT	**AEMT**	**Buckeye**
Silence # ð	77% stop-like	79% stop-like	45% stop-like
Stop consonant # ð	73%	88%	33%
Fricative/affricate # ð	28%	33%	9%
Liquid # ð	16%	16%	8%
Vowel # ð	11%	18%	7%

Table 4.10: Stop-like /ð/ occurrence across segmental contexts for different databases

The occurrence rates observed from the read speech in TIMIT and the task-oriented spontaneous speech in AEMT are very similar. The biggest difference between the two is for contexts where stop consonants precede /ð/, with a 15% distinction. AEMT speakers need to be deliberate and clear in their instructions in order for their counterparts to efficiently and accurately navigate around the map. Thus, the speech in AEMT may be closer to read speech than non-task-oriented spontaneous speech.

The occurrence rates are significantly lower for the Buckeye corpus in every segmental context when compared to the other two databases; as much as a 55% difference for sequences where stop-consonants precede /ð/. Dialect differences may be one possible explanation for the low stop-like /ð/ occurrence in the Buckeye Corpus. Speakers in the Buckeye corpus are all Ohio residents, whereas speakers of the TIMIT and AEMT utterances are from various regions in the country. Furthermore, the speech in Buckeye is much more casual than the speech in TIMIT and AEMT. The high level of casualness in speech may be another factor leading to the low stop-like /ð/ occurrence in the Buckeye Corpus.

A trend seen in all three databases is the higher occurrence of stop-like /ð/ when the phoneme is preceded by a voiceless obstruent than a voiced obstruent. Correlation analyses suggest a weak relationship between the two. However, this does not rule out the possibility of voicing as a contributor to the stop-like modification of /ð/.

Contexts with a preceding /t/ often have the highest stop-like /ð/ occurrence rates. In addition, the absence of the /t/, /d/, and sometimes /k/ release when these consonants precede a stop-like /ð/ is observed in all three databases. This suggests that the articulators may have formed the stop-like-/ð/ closure before the release of the preceding stop consonant, thereby obliterating the release burst of the preceding stop consonant. Thus, there seems to be articulatory overlap when stop-like /ð/ is produced in contexts where the preceding phoneme is a stop consonant.

Implications of Data on Possible Causes of the Modification

The data suggest a number of possible factors that may contribute to the stop-like modification of /ð/. One factor may be the physiological mechanisms of speech production, where stop-like /ð/ is a result of the complex interplay of aerodynamic forces, articulatory movements and biological properties of the vocal apparatus. This possibility is supported by the finding that stop-like /ð/ may occur in any type of segmental or prosodic context. In addition, there are higher occurrence rates of the modification when /ð/ is preceded by an obstruent than by a vowel. This is not surprising since the transition from an obstruent to /ð/ requires articulators to move from one specific position to another despite whatever aerodynamic and biological conditions are present; whereas the transition from a vowel, with its relatively open vocal tract and no pressure build-up, to /ð/ is much less complicated.

The role of intra-oral pressure may be one possible contributor to the stop-like modification of /ð/. The presence of intra-oral pressure plays an important role in the maintenance of a constriction, since it applies downward force on the tongue surface, as well as force leading to some displacement of the vocal tract walls. Thus, a constriction area may tend to increase with increased intra-oral pressure (Badin, 1989; Stevens, 1998). In the absence of intra-oral pressure, the static area of the oral constriction may be zero or negative (pushed up further up against the palate).

In sequences where an obstruent consonant precedes /ð/, intra-oral pressure is built up during the preceding obstruent's period of closure or constriction. When the preceding consonant is released, the intra-oral pressure is also released (Stevens, 1998). Therefore, the necessary intra-oral pressure for maintaining a /ð/ constriction needs to build up after the preceding consonant's release. It is possible that the articulators transitioning from a preceding obstruent's release to /ð/ may reach the target position before sufficient intra-oral pressure is built up, leading to a collapse of the constriction into a complete closure.

The role of intra-oral pressure may also contribute to the high occurrence of stop-like /ð/ when the phoneme is preceded by silence, such as those in utterance-initial position. Lower subglottal pressure is typical at the beginning of an utterance (Slifka, 2003). This lower subglottal pressure would lead to lower intra-oral pressure as a constriction is made in the oral cavity. Thus, it is possible that not enough intra-oral pressure is present for the maintenance of an oral constriction. The investigation of the physiological aspects of stop-like /ð/ will be discussed in-depth in Chapter IX.

Physiological factors do not seem sufficient in explaining the notably lower occurrence of stop-like /ð/ in non-task-oriented spontaneous speech of Buckeye than in the read speech of TIMIT and task-oriented spontaneous speech of AEMT. Speakers produced stop-like /ð/ more often in speech that is expected to be clearer and more deliberate than in speech that is expected to be more casual. In addition to speaking style and manner, dialect differences may also be an influencing factor for the stop-like modification of /ð/.

Furthermore, an upward trend is observed in the prosodically-labeled AEMT recordings, where the stop-like occurrence increases with higher positions on the prosodic hierarchy. Within each database, a relatively high occurrence of stop-like /ð/ is also observed when the phoneme is preceded by silence, most often in utterance or phrase initial positions. These findings suggest that the stop-like modification of /ð/ may be a form of articulatory strengthening, as characterized by greater palato-lingual contact. This possibility has been suggested for Dholuo dentals /ð θ/, where the consonants displayed more stop-like characteristics in word-initial position than medial position (Degenshein, 2004). Thus, a next step in studying the role of prosody in the stop-like modification of /ð/ in American English is to compare /ð/ cases, with the same segmental context, across prosodic domains.

The high rate of stop-like /ð/ occurrence when the phoneme is preceded by a stop consonant is consistent with previous publications (Roach, 2000; Cao, 2002; Shockey, 2003). The strong connection between the stop-like modification and the presence of a preceding stop consonant suggests that assimilation of manner may be a possible explanation. This explanation seems consistent with the behavior of /ð/, which has been found to often take on the manner characteristics of the preceding consonant (Manuel, 1995; Manuel and Wyrick, 1999). However, this study finds that high occurrences of stop-like /ð/ are not exclusive to segmental contexts where the consonant is preceded by a stop consonant. In fact, stop-like /ð/ occurs, to varying degrees, in all types of segmental contexts. Thus, it is more likely that the mechanisms that underlie the stop-like modification of /ð/ are physiological and/or prosodic in nature. Future directions in the investigation of possible causes of the stop-like modification are discussed in Chapter IX.

Summary

This study has examined word-initial /ð/ in the read speech of TIMIT, the task-oriented spontaneous speech of AEMT, and the non-task-oriented spontaneous speech of Buckeye. The word-initial /ð/ is most often found in the function word "the" in all three databases. The most common segmental contexts of /ð/ include ones preceded by /t/ (e.g. "at the", "about the"), /r/ (e.g. "for the"), /v/ (e.g. "of the"), and /z/ (e.g. "is the") in the spontaneous speech of AEMT and Buckeye. The least common segmental contexts in the two databases include ones preceded by /b/, /g/, /j/, /ch/, /sh/ and /zh/.

Through the acoustical analysis of /ð/, it has been found that high occurrences of stop-like /ð/ is not exclusive to segmental contexts where the consonant is preceded by a stop consonant. The data indicate that stop-like /ð/ occurs most often when preceded by silence (77% in TIMIT, 79% in AEMT, and 45% in Buckeye) or when preceded by a stop consonant (73% in TIMIT, 88% in AEMT, and 33% in Buckeye). The occurrence is less frequent (28% in TIMIT, 33% in AEMT, and 9% in Buckeye) when /ð/ is preceded by a fricative or an affricate. This modification rarely occurs (14% in TIMIT, 16% in AEMT, and 7% in Buckeye) when /ð/ is preceded by a vowel or liquid consonant.

A comparison of the three databases shows that speakers produced stop-like /ð/ much more often in the read speech (TIMIT) and task-oriented spontaneous speech (AEMT) than in non-task-oriented spontaneous speech (Buckeye). This finding does not support the hypothesis that stop-like /ð/ frequency would increase with the level of casualness in speech. Furthermore, stop-like /ð/ occurrence is observed to increase with higher positions on the prosodic hierarchy in the prosodically-labeled utterances of AEMT.

A trend that is consistently present in all three databases is a higher rate of stop-like /ð/ occurrence when it is preceded by a voiceless stop than a voiced stop. The same trend is seen when comparing /ð/ preceded by a voiceless fricative to that preceded by a voiced fricative. However, correlation analyses indicate a very weak correlation between the voicing of the preceding obstruent and the stop-like modification of /ð/.

These findings suggest that the stop-like modification may have a number of contributing factors, including physiological mechanisms of speech production, prosody, dialect, and/or other aspects of speaking style and manner. The investigation of some of these possible contributors will be discussed in Chapter IX. The next chapter will present the detailed acoustic analysis and comparisons of stop-like /ð/ and /d/.

V. ACOUSTICS OF STOP-LIKE /ð/

This section explores the following questions:

- Is there acoustic evidence suggesting that the dental place of articulation is preserved for stop-like /ð/?
- Are there salient acoustic cues that distinguish between stop-like /ð/ and /d/?

Based on tube approximations and findings of previous studies, it is hypothesized that stop-like /ð/ will have acoustic evidence suggesting dental-place preservation despite modification. Stop-like /ð/ is expected to have higher burst peak frequency, lower burst amplitude, flatter burst amplitude shape, and lower F2 at following vowel onset than /d/. Thus, these acoustic cues should be salient in distinguishing stop-like /ð/ from /d/.

Research Methodology

Acoustic and spectral-moment measures are taken from 100 word-initial stop-like /ð/ tokens and 102 /d/ tokens. The tokens are obtained from TIMIT directories 2 (speakers from northern region), 3 (speakers from north midland region), 7 (speakers from western region) and 8 (speakers moved around the country). All cases are followed by an unstressed vowel, [ə] or [ɨ], and extracted from utterances of 23 female and 59 male speakers. Each speaker has at least one /ð/ case and one /d/ case included in this analysis. It is noted that the number of tokens for stop-like /ð/ and /d/ differ by two; this difference is not expected to affect the comparisons of means and variances between the two consonants across the acoustic and spectral measures.

Tables 5.1 and **5.2** list the segmental contexts of stop-like /ð/ and /d/ tokens, respectively. As the tables show, stop-like /ð/ tokens are most often preceded by silence, while /d/ cases are most often preceded by a vowel.

Context of /ð/	**Number of Cases**
/ð/ preceded by silence	*54*
/ð/ preceded by a stop consonant	*28*
t#	17
d#	8
k#	2
b#	1
/ð/ preceded by a fricative or an affricate	*14*
v#	4
s#	4
z#	2
ð#	2
j#	2
/ð/ preceded by [ə] or [ɨ]	*2*
/ð/ preceded by l#	*2*

Table 5.1: Segmental context of stop-like /ð/ cases, arranged by the number of cases. All /ð/ cases are word-initial. The '#' sign indicates that the phoneme precedes /ð/ across word boundary.

The difficulty of finding /d/ in unstressed syllables has prompted the inclusion of cases that are not word-initial. Cases of non-nasalized /d/, with preceding nasal consonants, are also included. Overall, an attempt has been made to include a variety of segmental contexts for stop-like /ð/ and /d/ cases.

Context of /ð/	**Number of Cases**
/d/ preceded by [ə] or [ɨ]	*28*
Across word boundary	19
In the same word	9
/ð/ preceded by a liquid	*24*
r#	18
r-	2
l#	2
l-	2
/d/ preceded by a nasal	*22*
n#	13
ng#	5
n-	4
/d/ preceded by a fricative or an affricate	*15*
z#	8
s#	5
f#	1
v#	1
/ð/ preceded by a stop consonant	*8*
t#	7
d#	1
/d/ preceded by silence	*5*

Table 5.2: Segmental context of /d/ cases. Not all /d/ cases are word-initial. The '-' sign indicates that the phoneme precedes /d/ within the same word. Whereas the '#' sign indicates that the phoneme precedes /d/ across word boundary.

Acoustic measurements are done using MATLAB. Discrete Fourier Transforms (DFTs) are calculated with a Hamming window of 6.5ms and a sampling rate of 16 kHz, starting from the beginning of the burst to the end of the burst, with 1ms shifts. An average burst spectrum is taken from the DFTs calculated for each /ð/ and /d/ token. Formant and other vowel spectral measurements are made from first glottal period at following-vowel onset.

Acoustic Measures

The high-frequency spectrum peak location (F_{hi}) is measured in the consonant segment. As illustrated in **Figure 5.1**, this variable is measured by first dividing the averaged burst spectrum into eight 1-kHz frequency bands. The burst peak is chosen as the frequency with the highest amplitude in the band with the highest overall energy, above 2-kHz. The approximate dividing point is chosen to be 2-kHz in order to exclude neck-tissue resonance and back cavity resonances excited by vocal fold vibration, or back cavity resonances excited by aspiration if voicing is absent; some low-frequency energies contributed by the front-cavity noise source may also be present due to the coupling of the front and back cavities (Stevens, 1998). Since both stop-like /ð/ and /d/ are voiced, the low-frequency regions of the spectrum are expected to be similar. However, F_{hi} is expected to be higher for the dental place of articulation, which involves a shorter front cavity, than for the alveolar place.

It is noted that the spectrum is calculated using a short-time window of 6.5ms; frequency components within 300-500Hz of one another may not be easily distinguished. However, the difference in the burst peak of a dental and an alveolar is expected to be on the order of thousands of Hertz. The main front-cavity resonance of a dental obstruent is around 7 to 8 kHz, while that of an alveolar obstruent is around 4 to 5 kHz (Jongman, Wayland, and Wong, 2000; Stevens, 1998). Thus, this measure should be an adequate approximation of the burst peak frequency differences between the two consonants. Furthermore, an additional measure of the mean frequency in the burst spectrum above 2-kHz is calculated; that measure is discussed later on in this section.

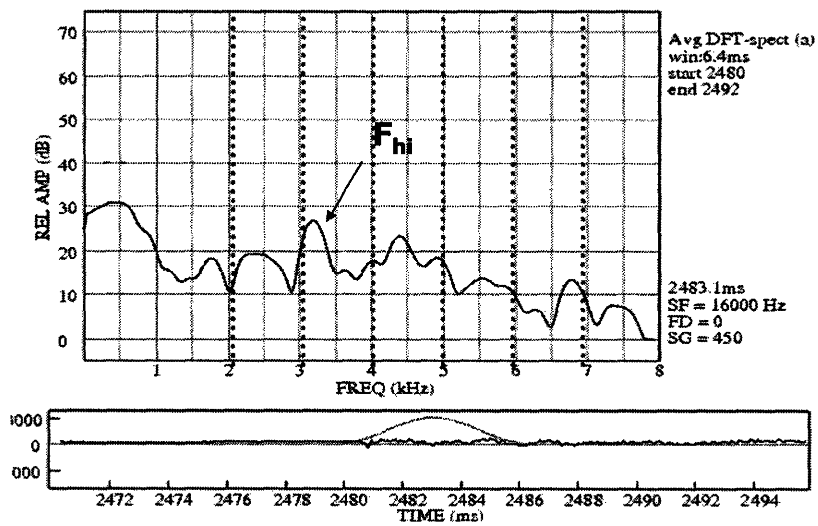


Figure 5.1: Burst spectrum peak frequency (F_{hi}). In the averaged spectrum, the burst peak is chosen as the frequency with the highest amplitude in the band with the highest overall energy, above 2-kHz. This is expected to be higher for the dental, which has a shorter front cavity, than for the alveolar.

The second cue measured is the ratio of the peak amplitude above 2-kHz to that below 2-kHz (A_{hi}-to-A_{lo}), shown in **Figure 5.2**. A_{hi} is the amplitude measured at F_{hi}. A_{lo} is the amplitude measured at the frequency with the highest amplitude in the band with the highest overall energy below 2-kHz. Again, the approximate dividing point is chosen to be 2-kHz, in order to separate low-frequency resonances of the back cavity and/or neck-tissues from the high-frequency resonances of the front cavity. Since both stop-like /ð/ and /d/ are voiced, the low-frequency regions of the spectrum are expected to be similar. Thus, they serve as good reference points to normalize the high-frequency regions of the spectrum. This variable is expected to be lower for dental consonants, which have lower energies in high frequencies than alveolar consonants.

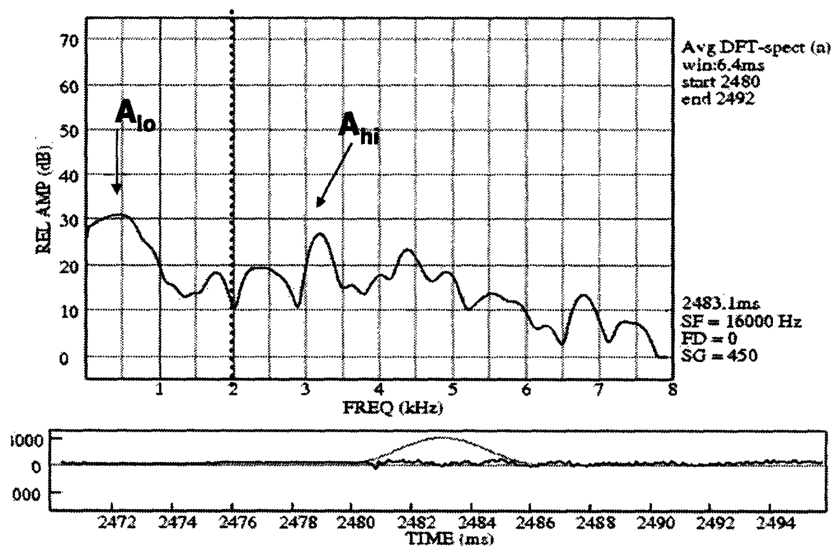


Figure 5.2: Ratio of peak amplitude above 2 kHz to that below 2 kHz (A_{hi}-to-A_{lo}), taken in the averaged spectrum. The ratio is expected to be lower for the dental, which has lower energies in high frequencies, than for the alveolar.

As illustrated in **Figure 5.3**, the ratio of total amplitude above 2-kHz in the consonant to that in the following vowel (C-to-V) is also measured. This variable is expected to be smaller for dentals due to their lower amplitude in high frequencies than for an alveolar. Again, the spectral region below 2-kHz is not considered in order to exclude low-frequency energies.

The fourth cue is the second formant frequency (F2) at following vowel onset, illustrated in **Figure 5.4**. This cue is expected to be lower for the dental, with its longer back cavity and more backed tongue body, than for the alveolar.

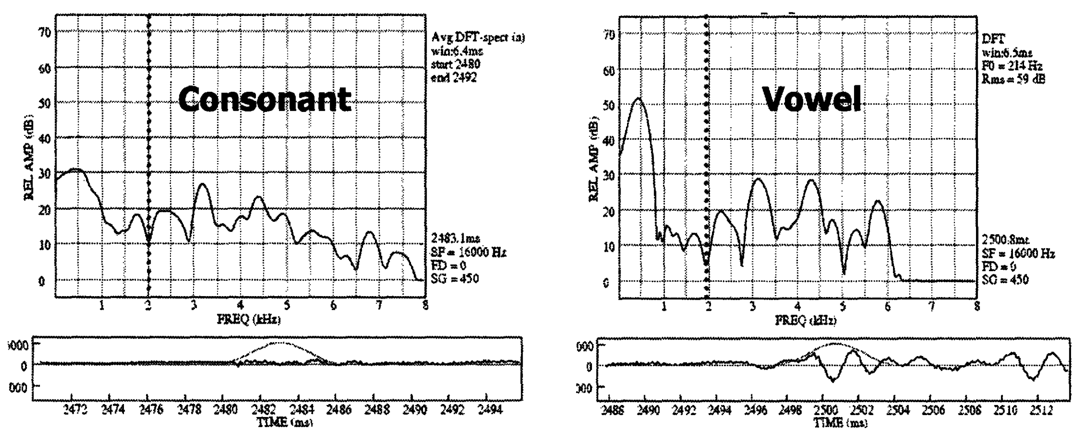


Figure 5.3: Ratio of total energy above 2 kHz in consonant to that in the vowel (C-to-V). This ratio is expected to be lower for the dental, which has lower energies in high frequencies, than for the alveolar.

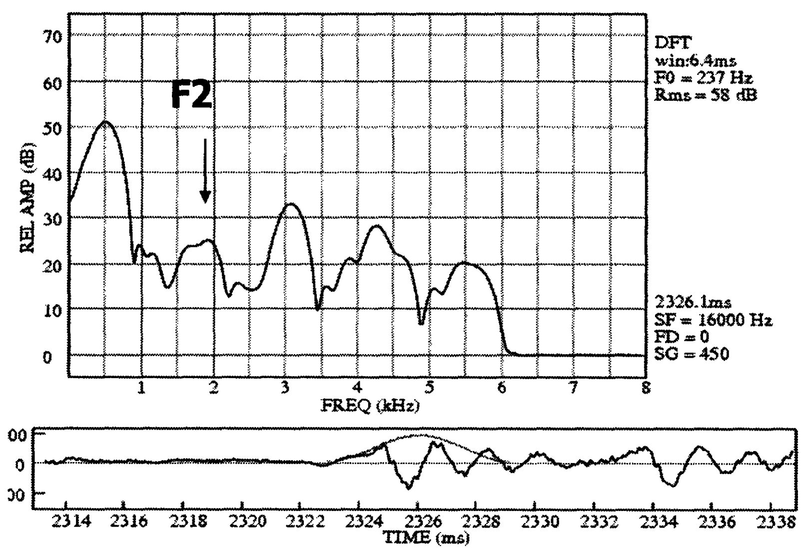


Figure 5.4: F2 at following vowel onset. This value is expected to be lower for the dental, which has a more back tongue body and longer back cavity, than for the alveolar.

Table 5.3 lists the expected differences between a dental and an alveolar stop for the four acoustic cues mentioned. It is noted that the differences are relative between dental and alveolar obstruents. Quantitative thresholds are not used to differentiate the two obstruents because there is high variability in speech among different individuals and different environments.

Acoustic Cue	**Expected Difference**	**Reason**
Burst Spectrum Peak (F_{hi})	Higher for Dental	Dentals have a shorter front cavity than alveolars
Peak amplitude above 2 kHz to that below 2 kHz (A_{hi}-to-A_{lo})	Lower for Dental	Dentals have a noise source that is both weaker and in a worse location for exciting the shorter front cavity than for alveolars
Total energy above 2 kHz in the consonant to that in the following vowel (C-to-V)	Lower for Dental	
F2 at following vowel onset	Lower for Dental	Dentals have a more backed tongue body and a longer back cavity

Table 5.3: The expected acoustic differences between dental and alveolar obstruents.

In addition to these acoustic cues, spectral moments are used to obtain measures of energy distribution across frequencies above 2-kHz in the burst spectrum. Spectral moments were first used by Forrest and colleagues (1988) to analyze voiceless obstruent consonants. Since both stop-like /ð/ and /d/ are voiced, the spectral region below 2-kHz is not considered in the calculation of spectral moments for reasons previously mentioned in this chapter.

The formulas for the spectral moments are listed below (Forrest et al., 1988). The first spectral moment is the mean frequency or center of gravity of the spectrum. It is used to calculate the other spectral moments. The second spectral moment is the variance around the mean frequency; this moment is used to calculate the standard deviation of the spectrum. The third moment is the skewness of the spectrum. It is used with the second spectral moment to calculate the dimensionless coefficient of skewness. Negative skewness indicates that there is more data on the left side of the spectrum than what is expected for a normal distribution. Positive skewness indicates that there is more data on the right side of the spectrum than what is expected for a normal distribution. The fourth moment is the kurtosis of the spectrum. It is used with the second moment to calculate the dimensionless coefficient of kurtosis, a measure of the peakedness of the spectrum. A factor of 3 is a correction to make the kurtosis of the normal distribution equal to zero.

Let $p(k) = \frac{amplitude(k)}{amplitude_{total}}$

k = sample number; f_k = frequency at kth sample; K= total number of samples

First Spectral Moment (L1): $$L_1 = \sum_{k=1}^{K} f_k \, p(k) \qquad \text{Eq.1}$$

Second Spectral Moment (L2): $$L_2 = \sum_{k=1}^{K} (f_k - L_1)^2 \, p(k) \qquad \text{Eq.2}$$

Third Spectral Moment (L3): $$L_3 = \sum_{k=1}^{K} (f_k - L_1)^3 \, p(k) \qquad \text{Eq.3}$$

$$Coefficient\ of\ Skewness = \frac{L_3}{L_2^{(3/2)}} \qquad \text{Eq.4}$$

Fourth Spectral Moment (L4): $$L_4 = \sum_{k=1}^{K} (f_k - L_1)^4 \, p(k) \qquad \text{Eq.5}$$

$$Coefficient\ of\ Kurtosis = \left(\frac{L_4}{{L_2}^2}\right) - 3$$ Eq.6

Figure 5.5 contains two spectra with different characteristics and consequently different spectral moment values. **Figure 5.5(a)** is obtained from the release burst of a /d/ token; **Figure 5.5(b)** is obtained from the voiced frication of a /z/ token. At regions above 2 kHz, **Figure 5.5(a)** has spectral peaks at around 3200Hz and 4400Hz, while **Figure 5.5(b)** has spectral peaks at around 4700Hz and 6500Hz. The general shape of the spectrum in **(a)** is flatter and broader than that of the spectrum in **(b)**. These differences are reflected in spectrum **(a)** having a lower mean frequency above 2 kHz (4200Hz) and larger standard deviation around the mean (1420Hz) than those of spectrum **(b)**, where the mean frequency is 5800Hz and the standard deviation is 1180Hz.

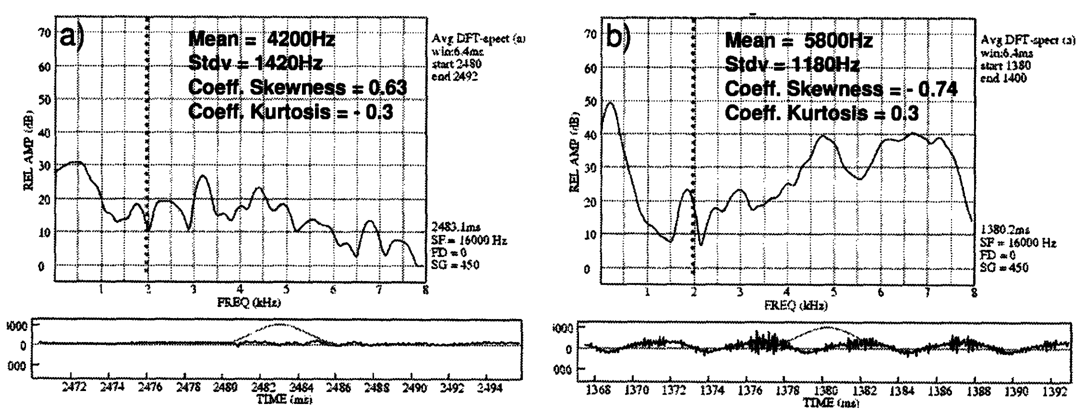


Figure 5.5: Spectra with different spectral moment values. Spectrogram in (a) is taken from a /d/ burst and the one in (b) is taken from /z/. Spectral moments are calculated for regions above 2 kHz in each spectrum. The spectrum on the left(a) is broader and flatter in shape than the spectrum on the right(b). In addition, (a) has lower-frequency peaks and more energy on the left side of the spectrum than (b). These differences are reflected in (a) having lower mean frequency, larger standard deviation, larger skewness, and smaller kurtosis than (b).

Furthermore, the coefficient of skewness is 0.63 for spectrum **(a)**, indicating more data in the left-half of the spectrum than what is expected for a normal distribution, while it is -0.74 for spectrum **(b)**, indicating more data in the right half of the spectrum than what is expected for a normal distribution. The coefficient of kurtosis, reflective of the peakedness of the spectrum, is lower for spectrum **(a)** than for spectrum **(b)**.

Table 5.4 lists the expected differences between the dental and alveolar stop in terms of spectral moments. The mean frequency above 2 kHz is expected to be higher for the dental than for an alveolar stop, due to the higher front cavity resonance for the dental than for the alveolar. The standard deviation around the mean frequency above 2 kHz is expected to be larger for a dental because of its flatter spectrum shape than that of an alveolar. Both the coefficient of skewness and coefficient of kurtosis are expected to be smaller for a dental, which has lower amplitude and flatter spectral shape in high frequencies, than an alveolar. It is noted that the differences are relative between dental and alveolar obstruents. Quantitative thresholds are not used to differentiate the two obstruents because there is high variability in speech among different individuals and different environments.

<table>
<tr><th>Spectral Measure</th><th>Expected Difference</th><th>Reason</th></tr>
<tr><td>Mean Frequency above 2kHz</td><td>Higher for Dental</td><td>Dentals have a shorter front cavity than alveolars</td></tr>
<tr><td>Standard Deviation (around Mean Frequency)</td><td>Larger for Dental</td><td rowspan="3">Dentals have a noise source that is both weaker and in a worse location for exciting the shorter front cavity than for alveolars. These factors lead to a lower amplitude and flatter spectral shape for dentals.</td></tr>
<tr><td>Coefficient of Skewness</td><td>Smaller for Dental</td></tr>
<tr><td>Coefficient of Kurtosis</td><td>Smaller for Dental</td></tr>
</table>

Table 5.4: Expected differences in spectral moments for dental and alveolar obstruents.

Results

Figure 5.6 shows the burst peak amplitude location (F_{hi}) for the 102 cases of /d/ ([d] in figure) and 100 cases of stop-like /ð/ ([dh] in figure). The median F_{hi} is 3770Hz for /d/ and 4330Hz for stop-like /ð/. There is considerable overlap in the data distributions between the two consonants. F_{hi} for /d/ ranges from 2870Hz at the 25th percentile to 4380Hz at the 75th percentile. F_{hi} for stop-like /ð/ ranges from 2740Hz at the 25th percentile to 5720Hz at the 75th percentile. Despite the overlap, the differences between the two sets of F_{hi} data are statistically significant, with a p-value smaller than 0.001. On average, F_{hi} is higher for stop-like /ð/ than for /d/. As listed in **Table 5.1**, this difference is expected if the dental place were preserved for the modified /ð/.

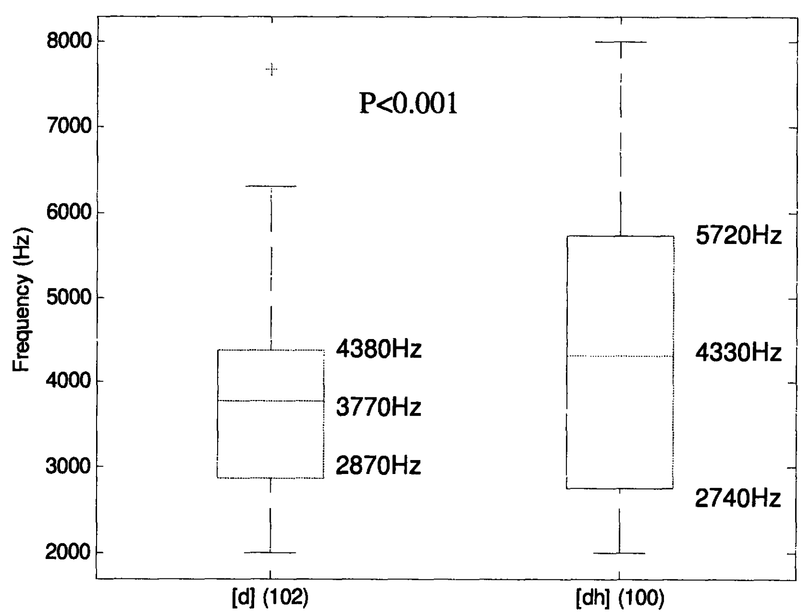


Figure 5.6: Burst spectrum peak (F_{hi}) for /d/ ([d] in figure) and stop-like /ð/ ([dh] in figure). Labeled in each box-plot, from top to bottom, are the 75th percentile, the median, and the 25th percentile, respectively.

Data on the ratio of the peak amplitude above 2 kHz and that below 2 kHz (A_{hi}-to-A_{lo}) for /d/ and stop-like /ð/ are shown in **Figure 5.7**. The median A_{hi}-to-A_{lo} is 1.9dB for /d/ and 0.02dB for stop-like /ð/. There is overlap in the data distributions for the two consonants. A_{hi}-to-A_{lo} for /d/ ranges from -0.28dB at the 25^{th} percentile to 4.4dB at the 75^{th} percentile. As for stop-like /ð/, A_{hi}-to-A_{lo} ranges from -2.2dB at the 25^{th} percentile to 1.9dB at the 75^{th} percentile. The differences between the two sets of data are statistically significant, with a p-value much smaller than 0.001. On average, A_{hi}-to-A_{lo} is smaller for stop-like /ð/ than for /d/. This difference is expected if the dental place were preserved for the modified /ð/ because dentals have lower and flatter amplitudes at high frequencies than alveolars.

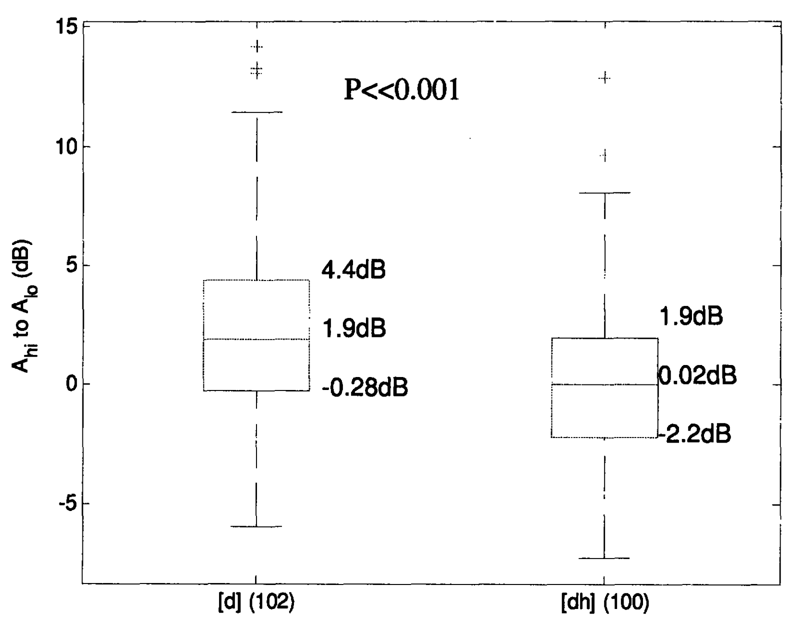


Figure 5.7: Peak amplitude above 2kHz to that below (A_{hi}-to-A_{lo}) in burst spectrum of /d/ ([d] in figure) and stop-like /ð/ ([dh] in figure) cases.

Figure 5.8 shows the total amplitude above 2kHz in the consonant to that in the following vowel (C-to-V) for /d/ and stop-like /ð/. The median C-to-V is 4.1dB for /d/ and 2.9dB for stop-like /ð/. The data for /d/ ranges from 0.3dB at the 25th percentile to 6.9dB at the 75th percentile; that for stop-like /ð/ ranges from -1.8dB at the 25th percentile to 7.1dB at the 75th percentile. As seen in the figure, the distribution for /d/ is more compact and falls within the distribution range for stop-like /ð/. The differences between the two sets of data are borderline statistically significant, with a p-value of less than 0.1. On average, C-to-V is higher for /d/ than for stop-like /ð/. This difference is expected if the dental place were preserved for the modified /ð/ because dentals have weaker energies in high frequencies than alveolars.

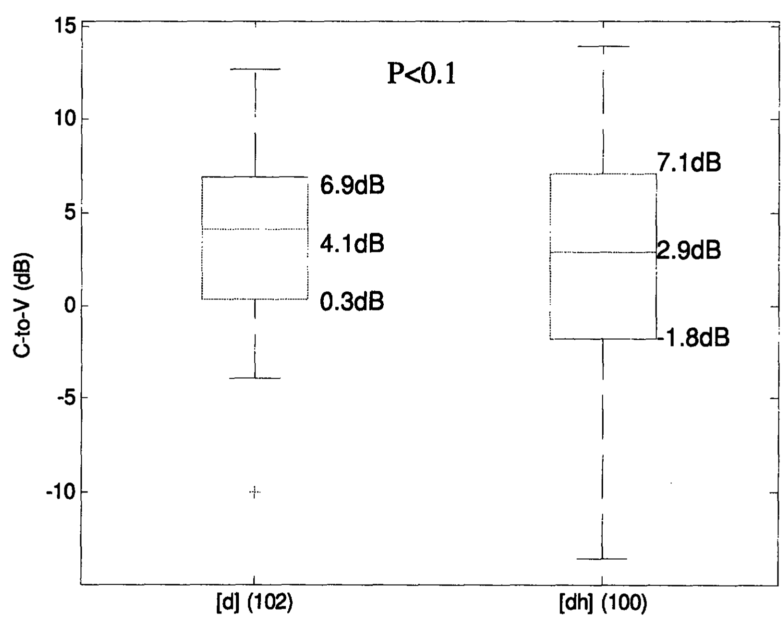


Figure 5.8: Total amplitude above 2kHz in consonant to that in the vowel (C-to-V) in burst spectrum of /d/ ([d] in figure) and stop-like /ð/ ([dh] in figure) cases.

F2 data taken at following vowel onset for /d/ and stop-like /ð/ are shown in **Figure 5.9**. The median is 1830Hz for /d/ and 1600Hz for stop-like /ð/. The distribution for /d/ ranges from 1600Hz at the 25th percentile to 2000Hz at the 75th percentile; that for stop-like /ð/ ranges from 1450Hz at the 25th percentile to 1850Hz at the 75th percentile. There is notable overlap between the two distributions. Nevertheless, F2 is on average higher for /d/ than for stop-like /ð/, with a statistically significant a p-value of less than 0.01. This difference is expected if the dental place were preserved for the modified /ð/ because dentals are produced with a longer back cavity and more backed tongue than alveolars.

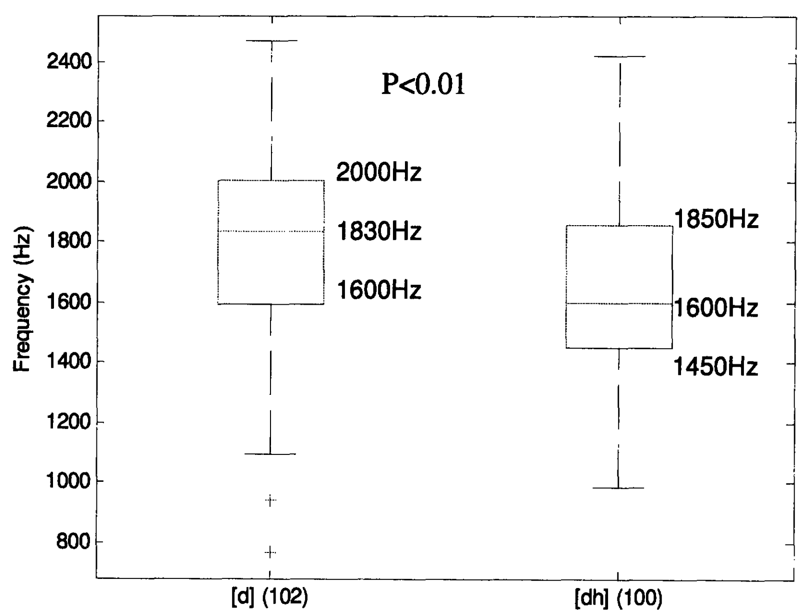


Figure 5.9: F2 at following vowel onset for /d/ ([d] in figure) and stop-like /ð/ ([dh] in figure).

Figure 5.10 contains data on the mean frequency of the averaged burst spectrum above 2kHz, for /d/ and stop-like /ð/ cases. The median is 4180Hz for /d/ and 4820Hz for stop-like /ð/. There is some overlap between the two distributions. The distribution for /d/ ranges from 3790Hz at the 25th percentile to 4470Hz at the 75th percentile; that for stop-like /ð/ ranges from 4540Hz at the 25th percentile to 5100Hz at the 75th percentile. The mean frequency is on average higher for stop-like /ð/ than for /d/; the difference is statistically significant, with a p-value of much less than 0.001. This difference is expected if the dental place were preserved for the modified /ð/ because dentals are produced with a shorter front cavity than alveolars.

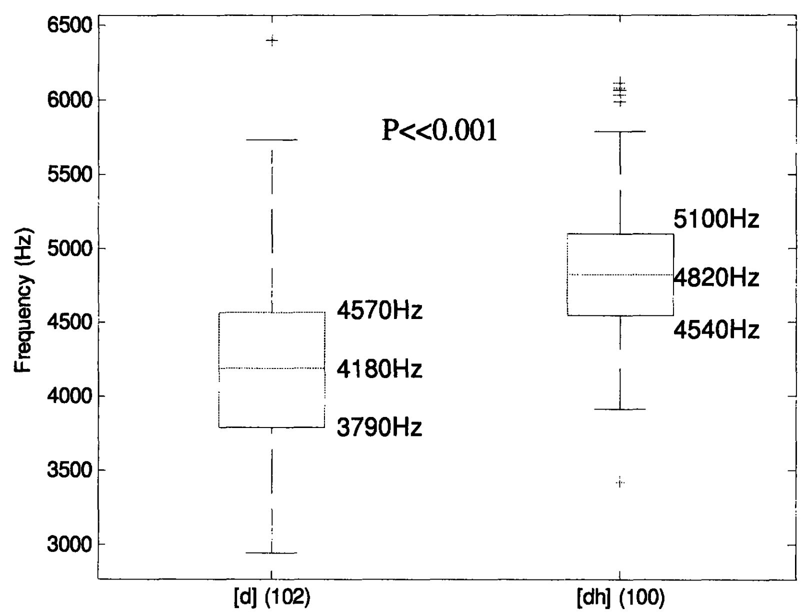


Figure 5.10: Mean frequency for /d/ ([d] in figure) and stop-like /ð/ ([dh] in figure).

Figure 5.11 shows the standard deviation around the mean frequency for /d/ and stop-like /ð/. The distribution for /d/ ranges from 1250Hz at the 25^{th} percentile to 1490Hz at the 75^{th} percentile; that for stop-like /ð/ ranges from 1530Hz at the 25^{th} percentile to 1800Hz at the 75^{th} percentile. The median is 1340Hz for /d/ and 1640Hz for stop-like /ð/. Despite some overlap between the two distributions, the standard deviation is on average larger for stop-like /ð/ than for /d/. Also, the difference is statistically significant, with a p-value of much less than 0.001. This difference is expected if the dental place were preserved for the modified /ð/. Dental stop consonants are expected to have large standard deviations around the mean frequency because they typically have flatter spectral shapes than those of alveolar stop consonants.

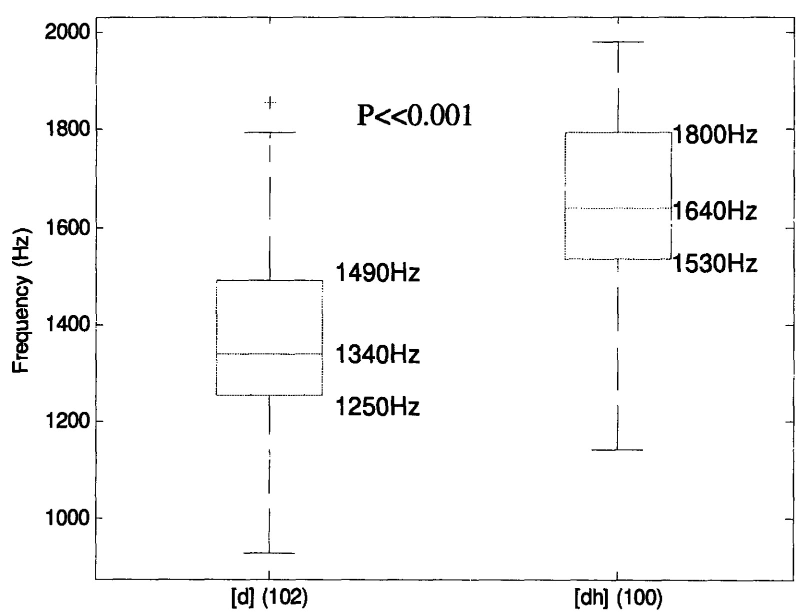


Figure 5.11: Standard deviation around mean frequency above 2kHz in the averaged burst spectrum for /d/ ([d] in figure) and stop-like /ð/ ([dh] in figure).

The coefficient of skewness, calculated for /d/ and stop-like /ð/, is graphed in **Figure 5.12**. The distribution is more compact for stop-like /ð/, with values ranging from -0.16 at the 25th percentile to 0.32 at the 75th percentile. As for /d/, the data distribution ranges from 0.3 at the 25th percentile to 1.2 at the 75th percentile. The median value for /d/, at 0.7, is higher that that for stop-like /ð/ at 0.1. On average, /d/ tokens have higher coefficient of skewness than stop-like /ð/ tokens. This indicates that /d/ tokens have more energy in the left side of the spectrum (2kHz to 5kHz) than stop-like /ð/ tokens. The front-cavity resonance of alveolar obstruents in around 4-5kHz, whereas that of dental obstruents is around 7-8kHz. Thus, this difference in skewness is expected if the dental place were preserved for the modified /ð/. Furthermore, the difference between the two sets of data is statistically significant, with a p-value much smaller than 0.001.

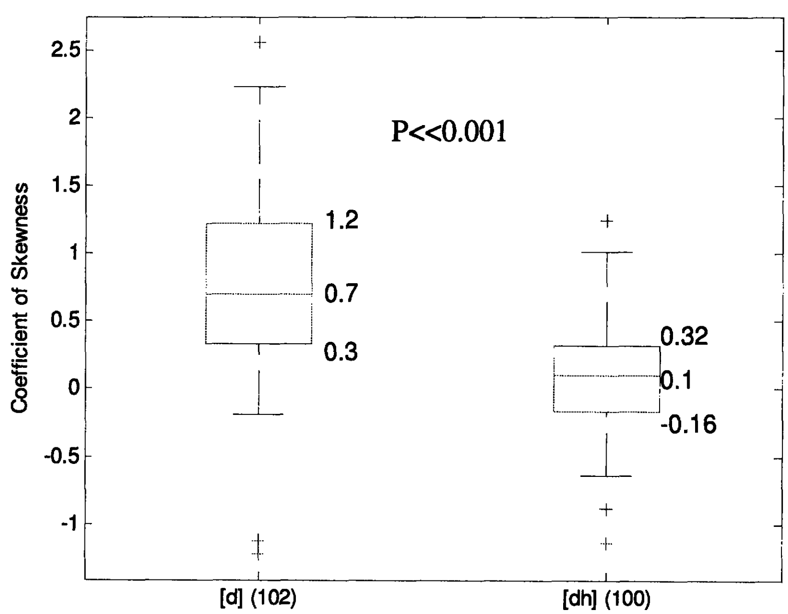


Figure 5.12: Coefficient of Skewness for /d/ ([d] in figure) and stop-like /ð/ ([dh] in figure).

Figure 5.13 contains data on the coefficient of kurtosis, calculated for the two consonants. The distribution for stop-like /ð/ is much more compact than that for /d/. Its coefficient of kurtosis ranges from -1.3 at the 25th percentile to -0.75 at the 75th percentile. On the other hand, the distribution is much larger for /d/, ranging from -0.55 at the 25th percentile to 1.4 at the 75th percentile. The median for /d/ is 0.19 while that for the stop-like /ð/ is -1.0. On average, the coefficient of kurtosis, which is a measure for the peakedness of the spectrum, is smaller for stop-like /ð/ than for /d/. This difference is expected if the dental place were preserved for the modified /ð/. Also, the difference between the two sets of data are statistically significant, with a p-value much smaller than 0.001.

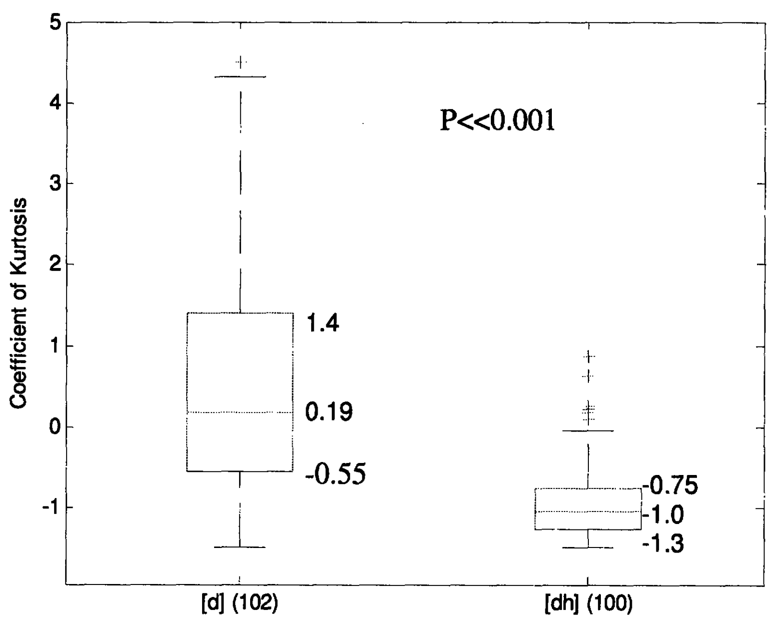


Figure 5.13: Coefficient of Kurtosis for /d/ ([d] in figure) and stop-like /ð/ ([dh] in figure).

Discussion

Table 5.5 is a summary of the acoustic and spectral moment analysis for the 102 /d/ tokens and 100 stop-like /ð/ tokens. The data reveal statistically significant differences between stop-like /ð/ and /d/ in all of the acoustic and spectral-moment measures analyzed.

Measure	/d/ (102 cases)	Stop-like /ð/ (100 cases)	P-Value	As expected if stop-like /ð/ has dental place?
A_{hi}-to-A_{lo}	Median = 1.9dB Mean = 2.5dB Stdv = 3.8dB	Median = 0.02dB Mean = 0.22dB Stdv = 3.3B	P<<0.001	Yes
F_{hi}	Median = 3770Hz Mean = 3730Hz Stdv = 1060Hz	Median = 4330Hz Mean = 4480Hz Stdv = 1830Hz	P<0.001	Yes
F2 at vowel onset	Median = 1830Hz Mean =1800Hz Stdv = 300Hz	Median = 1600Hz Mean = 1670Hz Stdv = 330Hz	P<0.01	Yes
C-to-V	Median = 4.1dB Mean = 3.9dB Stdv = 4.3dB	Median = 2.9dB Mean = 2.5dB Stdv = 6.2dB	P<0.1	Yes
Mean Frequency	Median = 4180Hz Mean =4200Hz Stdv = 630Hz	Median = 4820Hz Mean = 4860Hz Stdv = 480Hz	P<<0.001	Yes
Standard Deviation	Median = 1340Hz Mean = 1370Hz Stdv = 200Hz	Median = 1640Hz Mean = 1640Hz Stdv = 170Hz	P<<0.001	Yes
Coefficient of Skewness	Median = 0.7 Mean = 0.79 Stdv = 0.67	Median = 0.1 Mean = 0.08 Stdv = 0.38	P<<0.001	Yes
Coefficient of Kurtosis	Median = 0.19 Mean = 0.61 Stdv = 1.6	Median = -1.0 Mean = -0.91 Stdv = 0.49	P<<0.001	Yes

Table 5.5: Acoustic and spectral moment data for /d/ and stop-like /ð/.

The differences found between stop-like /ð/ and /d/ cases are as expected if the dental place were preserved for the modified /ð/. On average, stop-like /ð/ has higher F_{hi} and higher mean frequency in the spectrum above 2kHz than /d/. This difference is expected if stop-like /ð/ is produced with a dental place of articulation because dentals are produced with a shorter front cavity than alveolars. When the shorter front cavity is excited by a noise source, the concentration of burst amplitude should be located at a higher frequency for the dental than for the alveolar.

Stop-like /ð/ also has lower F2 at vowel onset than /d/. This difference is expected if the dental place were preserved for the modified /ð/, since the tongue body is typically lower and more backed for the dental than for the alveolar. Furthermore, the back cavity is slightly longer for the dental than for the alveolar. These factors contribute to lower F2, a back-cavity resonant frequency, going into and out of the dental consonant than of the alveolar.

Statistically significant differences are observed in the spectral amplitudes and shapes of the two consonants. On average, stop-like /ð/ has smaller A_{hi}-to-A_{lo}, smaller C-to-V, larger standard deviation around the mean frequency, smaller coefficient of kurtosis and smaller coefficient of skewness than /d/. These differences are expected if stop-like /ð/ has a dental place of articulation. Dental consonants have lower energies and flatter spectral shape in high frequencies than alveolars.

All measures exhibit overlaps in the data distributions for /d/ and stop-like /ð/. The cue that exhibits the most overlap is C-to-V, the ratio of total amplitude above 2kHz in the consonant to that in the following vowel; the distribution for /d/ falls within the distribution for stop-like /ð/. The differences in C-to-V between the two groups are borderline statistically significant, with a p-value of less than 0.1; the highest p-value of all the measures examined. A possible explanation of this smaller than expected difference in C-to-V is that most of the /d/ tokens come from unstressed or reduced parts of words such as "departure" and "distress". Therefore, such /d/ cases may have lower overall amplitude than what is expected for unreduced cases. This factor may also contribute to the overlaps in distributions observed for other measures; in particular those affected by the consonant's noise-source strength, such as A_{hi}-to-A_{lo}, standard deviation around the mean frequency in the spectrum, coefficient of skewness and coefficient of kurtosis.

The burst peak location (F_{hi}) is another cue where the distribution for /d/ falls within the distribution for stop-like /ð/. Individual differences in vocal tract characteristics, especially the length, would likely contribute to the variation of F_{hi}. In addition, /ð/ is produced with a very short front cavity and weak noise source. Therefore, the amplitude(s) of the front cavity resonance(s) would be small due to the weak excitation and increased radiation losses at high frequencies (Stevens, 1989). The resulting relatively-flat shape of the stop-like /ð/ burst spectrum would be a factor contributing to the wide distribution of F_{hi} for stop-like /ð/.

Differences in vocal tract length and tongue shape among individuals may also contribute to data variations for measures affected by vocal tract resonances; such measures include F2 at following vowel onset and mean frequency in the burst spectrum in addition to F_{hi}.

The results of this analysis are consistent with those from previous studies, where dental stops are found to have smaller burst amplitude, smaller coefficients of skewness and coefficients of kurtosis than alveolar stops (Jongman, Blumstein, and Lahiri, 1985; Stoel-Gammon et al, 1994; Sundara, 2005). However, this study finds higher mean frequency above 2kHz in the burst spectrum for stop-like /ð/ (median=4820Hz) than for /d/ (median=4180Hz); this finding seems to conflict with the lower mean frequency found for the dental Canadian-French /d/ (median≈3200Hz for males; median≈2900Hz for females, as shown in Sundara 2005, Figure 4) than for the alveolar Canadian-English /d/ (median≈3800Hz for males; median≈5000Hz for females, as shown in Sundara 2005, Figure 4) and in Sundara's study.

While the mean frequency for the American-English /d/ is between the medians of the male Canadian-English /d/ and female Canadian-English /d/, the mean frequency for stop-like /ð/ is around 1000Hz higher than that for the dental Canadian-French /d/. This discrepancy may be explained by the exclusion of spectral regions below 2kHz in this study, while Sundara uses the entire spectrum to obtain mean frequency. Dental stops are expected to have low energies in high frequencies. Thus, the inclusion of the low-frequency region, which contains energy due to voicing, aspiration if voicing is absent, and/or front-and-back-cavity coupling, would most likely lower the mean frequency of the spectrum. Therefore, mean frequencies must be obtained from the same regions of the spectrum in order to directly compare results from different studies.

Summary

Acoustic measures are obtained for 100 stop-like cases of word-initial /ð/ and for 102 /d/ cases from TIMIT. All cases are followed by an unstressed vowel, [ə] or [ɨ], and extracted from utterances of 23 female and 59 male speakers. Each speaker has at least one /ð/ case and one /d/ case included in this study.

On average, stop-like /ð/ is found to have higher F_{hi} and higher mean frequency in the spectrum above 2kHz than /d/, both with statistical significance. The differences are expected if the dental place were preserved for stop-like /ð/; dentals have higher front cavity resonances because they have a shorter front cavity length than that of alveolars. In addition, stop-like /ð/ is found to have lower F2 at vowel onset than /d/, with statistical significance. This difference is also expected if the dental place were preserved for the modified /ð/; dentals have a lower and more backed tongue body F2, as well as a slightly longer back cavity, than alveolars.

Statistically significant differences are also observed in the spectral amplitudes and shapes of the two consonants. On average, stop-like /ð/ has smaller A_{hi}-to-A_{lo}, smaller C-to-V, larger standard deviation around the mean frequency, smaller coefficient of kurtosis and smaller coefficient of skewness than /d/. These differences are expected if stop-like /ð/ has a dental place of articulation. Dental consonants have lower energies and flatter spectral shape in high frequencies than alveolars.

Overall, the analysis results support the hypothesis that the dental place of articulation is preserved for stop-like /ð/. In addition, data for each acoustic and spectral measure demonstrate statistically-significant differences between stop-like /ð/ and /d/. All measures seem salient in distinguishing those two consonants.

Consistent with previous research on other modifications of /ð/, this study finds acoustic evidence indicating that the dental place is preserved despite the change in the fricative manner of production. The following chapter will explore the robustness of the acoustic and spectral-moment measures in the automatic classification of stop-like /ð/ and /d/.

VI. AUTOMATIC CLASSIFICATION OF STOP-LIKE /ð/ AND /d/

This section explores the following questions:

- Are the acoustic and spectral-moment measures obtained from the consonant burst and the following vowel robust in the automatic classification of stop-like /ð/, distinguishing it from /d/?
- What are the applications of this research for automatic speech recognition systems?

It is hypothesized that the measures obtained from the consonant burst and the following vowel will be robust in automatically distinguishing stop-like /ð/ from /d/. In addition, these salient measures can be used in feature-based automatic speech recognition systems to effectively extract the features of /ð/ despite modification.

Research Methodology

The second study of this thesis suggests that burst spectrum peak frequency (F_{hi}), ratio of peak amplitude above 2kHz to that below 2kHz (A_{hi}-to-A_{lo}), ratio of consonant-to-vowel energy above 2kHz (C-to-V), F2 at following vowel onset, mean frequency in burst spectrum above 2kHz, standard deviation around that mean frequency, coefficient of skewness, and coefficient of kurtosis are salient in distinguishing stop-like /ð/ from /d/. Therefore, all eight measures are used in this automatic classification experiment. It is noted that these acoustic and spectral-moment measures are taken either in the averaged burst spectrum or at the first glottal period of the following vowel for stop-like /ð/ and /d/.

Data for each measure is collected from the same 100 word-initial stop-like /ð/ tokens and for 102 /d/ tokens analyzed in the previous study of this thesis. Again, these cases come from TIMIT directories 2 (speakers from northern region), 3 (speakers from north midland region), 7 (speakers from western region) and 8 (speakers moved around the country). All cases are followed by an unstressed vowel, [ə] or [ɨ], and extracted from utterances of 23 female and 59 male speakers. Each speaker has at least one /ð/ token and one /d/ token included in this experiment. **Tables 6.1** and **6.2** list the segmental contexts of stop-like /ð/ and /d/ tokens, respectively. As the tables show, stop-like /ð/ tokens are most often preceded by silence, while /d/ cases are most often preceded by a vowel.

Context of /ð/	**Number of Cases**
/ð/ preceded by silence	*54*
/ð/ preceded by a stop consonant	*28*
t#	17
d#	8
k#	2
b#	1
/ð/ preceded by a fricative or an affricate	*14*
v#	4
s#	4
z#	2
ð#	2
j#	2
/ð/ preceded by [ə] or [ɨ]	*2*
/ð/ preceded by l#	*2*

Table 6.1: Segmental context of stop-like /ð/ cases, arranged by the number of cases. All /ð/ cases are word-initial. The '#' sign indicates that the phoneme precedes /ð/ across word boundary.

The difficulty of finding /d/ in unstressed syllables has prompted the inclusion of cases that are not word-initial. Cases of non-nasalized /d/, with preceding nasal consonants, are also included. Overall, an attempt has been made to include a variety of segmental contexts for stop-like /ð/ and /d/ cases.

Context of /ð/	**Number of Cases**
/d/ preceded by [ə] or [ɨ]	*28*
Across word boundary	19
In the same word	9
/ð/ preceded by a liquid	*24*
r#	18
r-	2
l#	2
l-	2
/d/ preceded by a nasal	*22*
n#	13
ng#	5
n-	4
/d/ preceded by a fricative or an affricate	*15*
z#	8
s#	5
f#	1
v#	1
/ð/ preceded by a stop consonant	*8*
t#	7
d#	1
/d/ preceded by silence	*5*

Table 6.2: Segmental context of /d/ cases. Not all /d/ cases are word-initial. The '-' sign indicates that the phoneme precedes /d/ within the same word. Whereas the '#' sign indicates that the phoneme precedes /d/ across word boundary.

Leave-one-out classification, where one token serves as the unknown test and the rest as training data, is repeated for all /d/ and stop-like /ð/ tokens. Two different algorithms are used to automatically distinguish stop-like /ð/ from /d/. The first algorithm decorrelates the eight acoustic cues and spectral moments using principal component analysis (PCA). PCA decorrelates the variables of a data set by projecting the data on the eigen vectors of its own covariance matrix (Jolliffe, 1986; also see Appendix F for review of PCA). PCA also chooses a new coordinate system for the dataset such that the greatest variance by any projection of the dataset comes to lie on the first axis, called the first principal component, the second greatest variance on the second axis, and so on. This transformation is performed on the collected data across all measures, for all tokens of /d/ and stop-like /ð/. The assignment of unknown tokens is based on the Euclidean distance of each unknown test case to the centroids of the de-correlated stop-like /ð/ and /d/ training data (see Appendix I for MATLAB source code used in this automatic classification experiment).

The second classification scheme uses linear discriminant analysis, which involves a linear combination of features that best separate the two classes of objects, to determine the category of the unknown token (see Appendix G for review of Linear Discriminant Analysis). This transformation is done using MATLAB's **classify** function and selecting the linear discriminant analysis option (see Appendix I for Matlab source code used in this automatic classification experiment).

While both algorithms decorrelate the variables of the data, a major difference between the two is that LDA maximizes between-class variability and minimizes with-in class variability, while PCA does not. However, an advantage of PCA over LDA is that it can be used to reveal relationships among the variables of a data set.

Results

Component Analysis

Using PCA, eight principal components are obtained from the eight acoustic cues and spectral moment measures. As shown in **Table 6.3**, the first principal component (PC1) accounts for 42.9% of the variance in the data; the second principal component (PC2) accounts for 23.5%; the third component (PC3) for 13% of the variance; and the fourth component (PC4) accounts for 9.5% of the variance. All other principal components account for a total of 11.1% of the variance in the data.

The factor loadings for each component are also shown in the table. The factor loadings indicate the amount of contribution from each acoustic cue or spectral moment to each principal component. Higher loadings indicate greater contribution to the principal component. Highly correlated measures will typically have high factor loads in the same principal components.

The factor loading indicates that the largest contributors to PC1 are the four spectral moment measures and F_{hi}. It is not surprising that these measures are correlated. Each successive spectral moment is based on the first moment, the mean frequency of the spectrum. Both the mean frequency above 2kHz and F_{hi} in the burst spectrum are reflective of the front cavity resonance. Thus, PC1 seems to be governed by the front cavity resonance and any correlated measures in the burst spectrum.

	PC1 (42.9%)	PC2 (23.5%)	PC3 (13%)	PC4 (9.5%)	PC5 (5.7%)	PC6 (3.2%)	PC7 (1.8%)	PC8 (0.4%)
Frequency of Burst Spectrum Peak (F_{hi})	**0.33**	**0.43**	0.18	0.04	**0.65**	**0.4**	**0.31**	0.03
F2 at Following Vowel Onset	0.02	0.18	**0.89**	0.37	0.18	0.1	0.002	0.03
Ratio of A_{hi}-to-A_{lo} in Burst Spectrum	0.006	**0.56**	0.25	**0.54**	**0.48**	**0.31**	0.04	0.08
Ratio of C (consonant) to V (following vowel) above 2kHz	0.03	**0.49**	0.31	**0.69**	**0.41**	0.02	0.11	0.006
Mean Frequency above 2kHz in burst spectrum	**0.50**	0.16	0.004	0.18	0.05	**0.27**	**0.44**	**0.65**
Standard Deviation around mean frequency above 2kHz in spectrum	**0.39**	**0.39**	0.13	0.17	0.16	**0.68**	**0.39**	0.12
Coefficient of Skewness in spectrum above 2kHz	**0.51**	0.06	0.01	0.1	0.006	**0.44**	0.12	**0.71**
Coefficient of Kurtosis in spectrum above 2kHz	**0.46**	0.23	0.08	0.09	**0.36**	0.07	**0.73**	0.22

Table 6.3: Factor analysis using PCA. The high factor loads for each principal component (PC) is bolded.

A_{hi}-to-A_{lo}, A_{hi}-to-V_{hi}, and F_{hi} have the three highest factor loads for PC2. Those three measures are the acoustic cues taken from the burst spectrum. Standard deviation around mean frequency has the fourth highest factor load for PC2. Although this is a spectral moment, it is reflective of the flatness of the burst spectrum. Overall, PC2 is dominated by the acoustic cues taken from the burst spectrum.

PC3 is dominated by F2 at vowel onset, with a high factor loading of 0.89. It seems that this component is mostly influenced by F2, a back cavity resonance. A_{hi}-to-A_{lo} and A_{hi}-to-V_{hi} have the highest factor loadings for PC4. Thus PC4 seems mostly influenced by amplitude measures of the burst spectrum.

PC5 is similar to PC2 in that A_{hi}-to-A_{lo}, A_{hi}-to-V_{hi}, and F_{hi} have the three highest factor loads. The coefficient of kurtosis has the fourth highest factor load for PC5. Nevertheless, PC5 seems to be dominated by the acoustic cues taken from the burst spectrum. **Table 6.4** contains a generalized list of dominant contributors to each principle component.

Principal Component	Dominant Contributor(s)
PC1 (42.9%)	Front cavity resonance and correlated measures in the burst spectrum
PC2 (23.5%)	Acoustic cues from burst spectrum
PC3 (13%)	Back cavity resonance
PC4 (9.5%)	Amplitude measures from burst spectrum
PC5 (5.7%)	Acoustic cues from burst spectrum
PC6 (3.2%)	Non-generalizable list of acoustic and spectral measures (Standard deviation, skewness, F_{hi}, A_{hi}-to-A_{lo}, and mean frequency)
PC7 (1.8%)	Non-generalizable list of acoustic and spectral measures (Kurtosis, mean frequency, standard deviation, and F_{hi})
PC8 (0.4%)	Non-generalizable list of acoustic and spectral measures (Mean frequency and Skewness)

Table 6.4: Dominant contributors to each principal component

PC6, PC7, and PC8 account for less than 6% of the variance in the data. The factor loading data suggest non-generalizable contributors to each of those principal components. PC6 is dominated by the standard deviation around the mean, followed by skewness, F_{hi}, A_{hi}-to-A_{lo}, and mean frequency, respectively. PC7 is mainly influenced by kurtosis, followed by mean frequency, standard deviation, and F_{hi}. PC8 is dominated by mean frequency and skewness from the factor loading data.

Automatic Classification using PCA and LDA

Table 6.5 lists the results of the automatic classification experiments. Leave-one-out classification using the PCA-transformed data yields 83% accuracy in classifying the 100 stop-like /ð/ cases and 86% accuracy in classifying the 102 /d/ cases. The classification is based on the Euclidean distance of the unknown token from the centroids of the stop-like /ð/ and /d/ data. The feature space consists of the eight principal components obtained from all eight acoustic and spectral measures listed in **Table 6.3**.

	Stop-like /ð/ (100 tokens)	**/d/ (102 tokens)**
PCA	83% accurately classified	86% accurately classified
LDA	88% accurately classified	80% accurately classified

Table 6.5: Results of automatic classification of stop-like /ð/ and /d/. Leave-one-out classification is performed on PCA-transformed data, using a Euclidean distance metric. Leave-one-out classification is also performed using LDA.

Leave-one-out classification using LDA yields 88% accuracy in classifying the 100 stop-like /ð/ cases and 80% accuracy in classifying the 102 /d/ cases. The feature space is obtained from all eight acoustic and spectral measures.

The results indicate that LDA performs slightly better in classifying stop-like /ð/ cases (88% accuracy) than PCA (83% accuracy). However, PCA is slighter better in classifying /d/ cases (86% accuracy) than LDA (80% accuracy). Surprisingly, LDA, which maximizes between-group separation and minimizes within-group variation, does not demonstrate a clear advantage over PCA. Misclassified cases from each classification scheme are analyzed for further insights as to the causes of the performance errors and differences.

Misclassified Cases

Table 6.6 lists the misclassified /d/ cases and whether their data fall within one standard deviation from the means of each acoustic cue and spectral moment measured (see Appendix H for actual values of each misclassified case). Most of the /d/ cases (fourteen out of twenty-one) are misclassified by both classification algorithms; six cases are misclassified by LDA only; one case is misclassified by PCA only.

Out of the fourteen cases that are misclassified by both algorithms, only two (#1 and #10) have fewer than three measures where data values fall outside of plus/minus one standard-deviation (±1σ) from their respective means. Of the cases that are misclassified by LDA only, five out six (#4, #8, #11, #12, and #18) have fewer than three measures where data values fall outside of $\pm 1\sigma$ from their respective means. The case (#17) that is misclassified solely by PCA has two measures with values outside of $\pm 1\sigma$. These observations indicate that LDA misclassifies /d/ more often than PCA, even when tokens have values within $\pm 1\sigma$ for most measures.

Case #	Misclassified by:	Within one standard deviation from the mean?							
		F_{hi}	A_{hi}:A_2	C:V	F2	Cog	Stdv	Skew	Kurt
1	PCA and LDA	Y	Y	Y	**N-**	**N+**	Y	Y	Y
2	PCA and LDA	**N-**	Y	Y	Y	Y	**N+**	**N-**	**N-**
3	PCA and LDA	**N+**	**N-**	**N+**	Y	**N+**	**N+**	**N-**	**N-**
4	LDA	Y	Y	Y	Y	Y	**N+**	Y	Y
5	PCA and LDA	**N-**	**N-**	Y	Y	**N+**	**N+**	**N-**	**N-**
6	PCA and LDA	**N+**	**N+**	**N+**	Y	**N+**	Y	**N-**	Y
7	PCA and LDA	**N-**	Y	Y	**N+**	Y	**N+**	**N-**	**N-**
8	LDA	Y	Y	Y	Y	Y	**N+**	Y	Y
9	PCA and LDA	**N+**	Y	Y	Y	Y	**N+**	Y	**N-**
10	PCA and LDA	Y	Y	Y	Y	Y	**N+**	Y	Y
11	LDA	**N+**	Y	Y	Y	Y	Y	**N-**	Y
12	LDA	Y	Y	Y	Y	Y	**N+**	Y	Y
13	PCA and LDA	**N-**	Y	Y	**N-**	Y	**N+**	Y	Y
14	PCA and LDA	**N+**	**N-**	Y	Y	Y	**N+**	Y	Y
15	PCA and LDA	**N-**	Y	**N-**	Y	Y	**N+**	Y	Y
16	PCA and LDA	Y	**N-**	Y	**N+**	**N+**	**N+**	**N-**	**N-**
17	PCA	**N+**	Y	Y	Y	Y	Y	**N-**	Y
18	LDA	Y	Y	**N-**	Y	Y	Y	Y	Y
19	PCA and LDA	**N-**	**N-**	**N+**	**N-**	Y	Y	Y	Y
20	PCA and LDA	**N-**	Y	Y	**N-**	Y	**N+**	Y	Y
21	LDA	**N+**	**N+**	Y	Y	**N-**	Y	**N-**	Y

Table 6.6: Misclassified /d/ cases. 'Y' indicates that the value is within one standard deviation; 'N' indicates that the value is outside of one standard deviation from the mean. '-' indicates that the value is one standard deviation below the mean; '+' indicates that it is one standard deviation above the mean

The table also indicates that five misclassified /d/ cases, #4, #8, #10 #12, and #18, have only one measure where data values fall outside of $\pm 1\sigma$. Four of them, cases #4, #8, #10, and #12, have very high values for the standard deviation around the mean frequency in the burst spectrum (**'Stdv'** in **Table 6.6**). This indicates that the shapes of their burst spectra are much broader and closer to what is expected for stop-like /ð/ than for /d/. Case #18 has a particularly low C-to-V ratio, making it closer to what is expected for /ð/ than for /d/.

Cases #3 and #16 are the only two cases having four or more measures where data values fall outside of $\pm 1\sigma$ from their respective means. In addition, those values are all closer to those expected for a stop-like /ð/ than a /d/. Although the two cases (from "does" and "did") do not sound like stop-like /ð/ to the author when played in context, both are found to sound extremely reduced. Reduction can lead to low energies as well as a broad and flat spectral shape for the /d/-burst spectrum.

Table 6.7 lists the misclassified /ð/ cases and whether their data fall within one standard deviation from the means of each acoustic cue and each spectral moment (see Appendix H for actual values of each misclassified case). Similar to the misclassified /d/ cases, most of the /ð/ cases (eleven out of eighteen) are misclassified by both PCA and LDA; six cases are misclassified by PCA only; one case is misclassified by LDA only.

Out of the eleven cases that are misclassified by both algorithms, only five (#8, #10, #12, #13, and #18) have fewer than three measures where data values fall outside of $\pm 1\sigma$ from their respective means. Of the cases that are misclassified by PCA only, two out of six (#9 and #15) have fewer than three measures where data values fall outside of $\pm 1\sigma$ from the mean. The case (#1) that is misclassified solely by LDA has two measures with values outside of ±1σ from the mean. These observations indicate that PCA misclassifies stop-like /ð/ more often than PCA.

Case #	Misclassified by:	Within one standard deviation from the mean?							
		F_{hi}	A_{hi}:A_{lo}	C:V	F2	Cog	Stdv	Skew	Kurt
1	LDA	**N-**	Y	Y	**N+**	Y	Y	Y	Y
2	PCA	**N+**	Y	**N+**	Y	**N+**	Y	**N-**	**N+**
3	PCA	Y	**N+**	Y	Y	**N+**	**N-**	Y	Y
4	PCA	**N-**	Y	Y	**N-**	**N-**	**N+**	Y	**N-**
5	PCA and LDA	Y	Y	Y	**N+**	**N-**		**N+**	Y
6	PCA and LDA	Y	Y	Y	**N-**	**N-**	**N-**	**N+**	Y
7	PCA and LDA	Y	Y	Y	Y	**N-**	**N-**	**N+**	**N+**
8	PCA and LDA	Y	Y	**N+**	Y	Y	Y	**N+**	Y
9	PCA	Y	**N+**	Y	Y	Y	Y	Y	Y
10	PCA and LDA	Y	**N+**	Y	Y	Y	Y	Y	Y
11	PCA and LDA	**N+**	Y	**N-**	Y	**N-**	**N-**	**N+**	**N+**
12	PCA and LDA	Y	Y	Y	**N+**	Y	Y	Y	Y
13	PCA and LDA	Y	**N+**	Y	Y	Y	Y	Y	Y
14	PCA and LDA	Y	**N+**	**N+**	**N+**	**N-**	**N-**	**N+**	**N+**
15	PCA	Y	Y	Y	**N+**	Y	Y	Y	Y
16	PCA and LDA	Y	**N+**	**N+**	**N+**	**N+**	**N-**	**N-**	**N+**
17	PCA	Y	Y	**N+**	Y	Y	**N-**	**N-**	**N+**
18	PCA and LDA	Y	Y	Y	Y	**N-**	Y	**N+**	Y

Table 6.7: Misclassified stop-like /ð/ cases. 'Y' indicates that the value is within one standard deviation; 'N' indicates that the value is outside of one standard deviation from the mean. '-' indicates that the value is one standard deviation below the mean; '+' indicates that it is one standard deviation above the mean

All but four of the misclassified /ð/ cases have at least two measures where data values fall outside of ±1σ from the mean. Cases #9 and #10 have very high A_{hi}-to-A_{lo} values, while all of their other measures are within one standard deviation from the mean. This indicates that their burst spectra have much higher energies in frequencies above 2kHz than below 2kHz, a characteristic that is expected for alveolar obstruents than for dental obstruents. Cases #12 and #15 have very high F2 at following vowel onset values, while all of their other measures are within one standard deviation from the mean. It is most likely that their high F2 values have brought them closer to what is expected for /d/ than for /ð/, thereby causing the classification error.

Cases #14 and #16 are the only two cases having four or more measures where data values fall outside of $\pm 1\sigma$ from the mean; in addition, the values are all closer to those expected for /d/. The two cases do not sound out of the ordinary to the author when played in context. However, both are preceded by /t/ in the phrase "at the". It may be possible that /ð/ is produced with an alveolar place of articulation in the two cases. However, these two cases seem to be in the minority of the twenty-five stop-like /ð/ cases with preceding /t/ or /d/; the rest do not exhibit overwhelming evidence for alveolar place of articulation.

Overall, **Tables 6.6** and **6.7** indicate that misclassification may occur for tokens with values outside of at least one standard deviation from the mean, for one or more measures. However, having one or more measures with values outside of one standard deviation from the mean does not automatically lead to the misclassification of a particular token. It seems that misclassification occurs when one or more outlying data values pull the token closer to the expected values of the wrong consonant class. However, misclassification may also be influenced by other factors such as the weight of each measure, token-specific data values, and data transformation method(s) for decorrelation and feature-space manipulation.

Discussion

This experiment has found that F_{hi}, A_{hi}-to-A_{lo}, C-to-V, F2 at following vowel onset, mean frequency in burst spectrum above 2kHz, standard deviation around that mean frequency, coefficient of skewness, and coefficient of kurtosis, can be used to automatically classify stop-like /ð/ and /d/ cases in the same following-vowel onset, to achieve 80+% accuracies.

Of the two different algorithms used in this experiment, leave-one-out classification using the PCA has performed roughly as well as that using LDA. LDA performs only slightly better in classifying stop-like /ð/ cases (88% accuracy) than PCA (83% accuracy). PCA is actually slighter better in classifying /d/ cases (86% accuracy) than LDA (80% accuracy). It seems that maximizing between-class variability and minimizing with-in class variability do not lead to huge improvements in classification accuracies.

There may be several factors contributing to the classification errors of stop-like /ð/ and /d/. Both PCA and LDA employ linear transformations of the dataset; non-linear relationships among the acoustic and spectral measures may not be captured. More sophisticated classification algorithms, such as ones that capture both linear and non-linear relationships, may achieve better results. In addition, the size of the training data may be a possible constraint on the performance of the classification tests. The number of cases may not be large enough to effectively model the consonants; if that were the case, then the sample size per consonant needs to be increased. Furthermore, the eight measures employed in this automatic classification experiment may not provide all the necessary information to distinguish stop-like /ð/ from /d/.

It might be fruitful to administer perception tests to find out how well human listeners can distinguish between stop-like /ð/ and /d/ cases. This would provide a basis of comparison for machine classification of stop-like /ð/ and /d/. If humans are very accurate (i.e. at least 90% accurate) in distinguishing stop-like /ð/ from /d/ in consonant-vowel sequences, then it is possible that the acoustic cues in the consonant and following vowel segment provide much of the information for accurate classification. However, if humans perform poorly for such cases, but not when more contextual information is provided (i.e. embedded in an utterance), then acoustic cues obtained only from the consonant and following vowel segments may not be enough for accurate classification; additional contextual information may be needed for accurate classification of stop-like /ð/.

Nevertheless, the eight measures from the consonant segment and its following vowel (F_{hi}, A_{hi}-to-A_{lo}, C-to-V, F2 at following vowel onset, and the first four spectral moments of the burst spectrum in regions above 2kHz) used in this study seem to be robust in automatically classifying stop-like /ð/, distinguishing it from /d/. Thus, these measures can be used by automatic speech recognizers to distinguish /ð/ from /d/ despite a change in manner of production. This may be particularly useful for feature-based speech recognizers, where these measures can be used to extract the features of /ð/ despite modification.

In addition to the saliency of these measures, this experiment also reveals important relationships among the acoustic cues and spectral moments through the analysis of principal components and factor loadings. It is found that factors influencing the first five principal components (which account for roughly 95% of the variance in the data) can be generalized into various classes of acoustic and spectral measures. For example, the first principal component is dominated by the front cavity resonance and correlated measures in the burst spectrum. The second and fifth principal components are dominated by acoustic cues from the burst spectrum. The back cavity resonance is the major contributor to the third principal component. The fourth component is mostly influenced by amplitude measures from the burst spectrum.

These relationships can be used for speech recognizers to group acoustic and spectral-moment measures into larger classes. These classes may also be ranked according to their relative contributions to the variance in the data. Doing so would provide another level of sophistication to the speech recognition system. For example, suppose a system determines the place of articulation for a consonant by using sixteen measures from the consonant segment and two measures from the following vowel segment. If conflicting values are obtained from the consonant segment, information from the following vowel would not make a large impact when each measure is considered separately; this is because there are many more measures taken from the consonant than from the vowel. However, if the measures are grouped into two separate classes, one from the consonant and the other from the following vowel, then the information obtained from the vowel may play a much bigger role in resolving the conflicts in the consonant-class measures. This concept may be particularly useful for the acoustic modeling stage of speech recognition systems.

Summary

Automatic classification is performed on 100 stop-like /ð/ cases and 102 /d/ cases with similar following-vowel context. These cases are obtained from utterances of 23 female and 59 male speakers in the TIMIT database. Each speaker has at least one /ð/ case and one /d/ case included in this experiment.

Two different classification algorithms are employed. The first algorithm involves leave-one-out classification using the PCA-transformed data. The feature space consists of the eight principal components obtained from the salient measures analyzed in the second experiment of the thesis; they are F_{hi}, A_{hi}-to-A_{lo}, C-to-V, F2 at following vowel onset, mean frequency in burst spectrum above 2kHz, standard deviation around that mean frequency, coefficient of skewness, and coefficient of kurtosis. The classification decision is based on the Euclidean distance of the unknown token from the centroids of stop-like /ð/ and /d/ data. This algorithm has yielded 83% accuracy in classifying the 100 stop-like /ð/ cases and 86% accuracy in classifying the 102 /d/ cases.

The second algorithm involves leave-one-out classification using LDA; this method has yielded 88% accuracy in classifying the 100 stop-like /ð/ cases and 80% accuracy in classifying the 102 /d/ cases. The feature space is also obtained from F_{hi}, A_{hi}-to-A_{lo}, C-to-V, F2 at following vowel onset, mean frequency in burst spectrum above 2kHz, standard deviation around that mean frequency, coefficient of skewness, and coefficient of kurtosis.

The 80+% classification accuracies obtained in this experiment suggest that the eight acoustic and spectral moment measures obtained from the consonant segment and the following vowel are robust in automatically classifying stop-like /ð/ and /d/. Therefore, these measures can be used by automatic speech recognizers to correctly identify stop-like /ð/ despite the change in manner of production.

In addition, this experiment has revealed important relationships among the acoustic cues and spectral moments through the analysis of principal components and factor loadings. It is found that factors influencing the first five principal components (which account for roughly 95% of the variance in the data) can be generalized into various classes of acoustic and spectral measures. These relationships can be used by speech recognizers to group acoustic and spectral measures into larger classes. These classes may be ranked according to their relative contributions to the variance in the data. Doing so would allow information from one or more classes to resolve conflicts or lack of information in another class(es), regardless of the number of measures in each class.

The next chapter will further discuss the application of these thesis findings in automatic speech recognition.

VII. APPLICATIONS IN AUTOMATIC SPEECH RECOGNITION

This thesis research has uncovered information on the context and acoustics of stop-like /ð/, a common variation of a frequently-occurring sound in spoken English. It has been revealed that stop-like /ð/ occurs more often in some contexts, such as ones with preceding silence or a preceding stop consonant, than in others. Acoustic evidence has indicated that despite the change in production manner, the dental place of articulation is preserved for /ð/. In addition, acoustic measures taken from the consonant burst spectrum and the onset of its following vowel seem robust in the automatic classification of stop-like /ð/ and /d/. Where does all this information fit in terms of applications in automatic speech recognition (ASR) systems? This chapter will first give an overview of the statistical speech recognition approach used in most commercial ASR systems. Next, the areas where the results of this thesis research may be applied will be highlighted. Applications of the thesis results in feature-based ASR systems will also be discussed.

Statistical ASR Systems

Most commercial ASR systems employ methods of statistical pattern matching to recognize speech (Young, 1996; Young, 2001; Glass and Zue, 2005). The basics of the statistical approach were pioneered in the 1970s by Jelinek and colleagues at IBM. **Figure 7.1** illustrates the basic steps in statistical speech recognition. The general method involves the conversion of the unknown speech signal into a sequence of acoustic vectors, $\mathbf{A} = \mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, ..., \mathbf{a}_T$. Each vector contains representations of the speech signal covering a certain period of time, typically 10ms. From there, the ASR system determines the most probable word sequence $\hat{\mathbf{W}}$, given the acoustic signal representation **A**. Using Bayes' Rule, that probability can be broken down into:

$$\hat{\mathbf{W}} = \arg\max_W P(W|A) = \arg\max_W \frac{P(W)P(A\,|\,W)}{P(A)} \qquad \text{Eq. 7}$$

Thus, the word sequence that maximizes the product of **P(W)** and **P(A|W)** needs to be found in order to estimate the most likely word sequence $\hat{\mathbf{W}}$. It is noted that 'arg max' stands for the value of the given argument that maximizes the value of the expression.

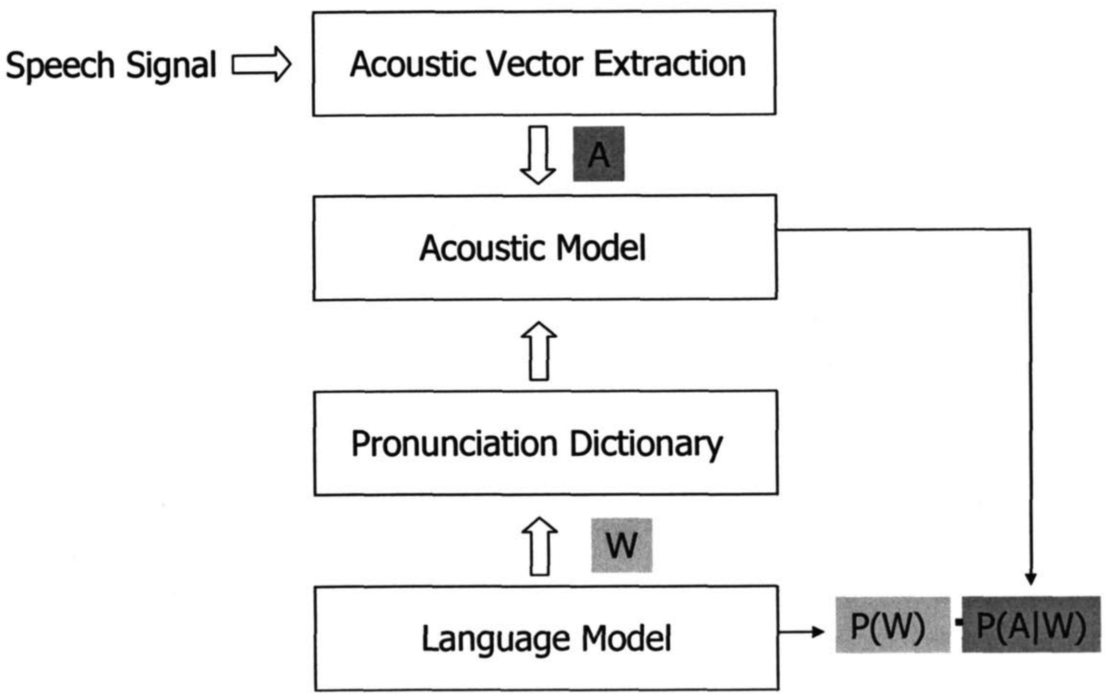


Figure 7.1: Basic steps in statistical speech recognition. The speech signal is represented by the vector sequence **A**, after acoustic vector extraction. The postulated word sequence **W**, from the language model, gets converted into basic speech units by the pronunciation dictionary. The probability of generating **A** given **W**, **P(A|W)**, is calculated using the acoustic model. The language model computes **P(W)**, the probability of observing **W** independent of the acoustic signal. The most likely word sequence is the one that maximizes **P(W)P(A|W)**.

The term **P(W)** is the *a priori* probability of observing the word sequence **W**, independent of the observed acoustic signal. This probability is determined by the language model, which is trained on data appropriate for the intended task of the ASR system. For example, if the ASR system is intended to recognize weather inquiries, then the language model will be trained on a corpus of such types of inquires. Based on the training data, the language model provides word predictions depending on the preceding words, in order to find the most likely word sequence.

The term, **P(A|W)**, is the probability of observing the acoustic sequence **A**, given the word sequence **W**. This probability is determined by the acoustic model, which is also trained on task-appropriate speech data.

As an example of the process, suppose a word sequence **W** = "The seat" is postulated. The probability of that sequence, **P(W)**, is computed by the language model. Then, a pronunciation dictionary is used to convert the words into a series of basic speech units; the most often used units are phones. A statistical model, most often a Hidden-Markov Model (HMM), is associated with each different phone. The probability of the particular sequence of HMMs needed to generate the observed acoustic sequence **A**, given the postulated word sequence **W**, is **P(A|W)**. This process can be repeated for all possible word sequences. The recognizer output is the most probable sequence.

Applications in Statistical ASR Systems

The results of this thesis research can be applied in two main areas of statistical ASR systems. The first area is in the extraction of acoustic vectors from the speech waveform. This research indicates that salient information, such as the spectral peak, spectral shape, and formant data, exists in the release burst of stop-like /ð/ and in the first glottal period of the following vowel; these measures distinguish stop-like /ð/ from /d/. Therefore, acoustic information can be used to resolve conflicts in utterances such as "the seat is dirty" and "deceit is dirty", when the language model estimates roughly equal probability for both. ASR systems that employ spectral representations of the acoustic signal can make sure that these measures are extracted from the speech signal in order to better capture the acoustic differences that differentiate between stop-like /ð/ and /d/.

The second application is in the area of acoustic modeling. The acoustic models (i.e. HMMs) need to have accurate representations of basic-acoustic-unit (i.e. phone) distribution in all types of contexts. These representations are obtained from the training data. Thus, sparse data most likely leads to poor representations. Furthermore, it is virtually impossible to cover all possible contexts through training alone. Since the system is probabilistic, poor representations will increase the error rate of the speech recognizer.

The frequency of stop-like /ð/ in various segmental contexts, as found in this study, is useful in that it provides distributions of stop-like /ð/ and non-stop-like /ð/ in a variety of contexts. This can be used in acoustic modeling to defend against a possibly sparse training set. For example, prior to training, the acoustic model may be initialized with high distributions of stop-like /ð/ in contexts where it is preceded by silence or by a stop consonant. On the other hand, that distribution may initially be low when the preceding phone is a vowel or sonorant consonant. These distributions may be modified as information is gathered from the training data; otherwise, they can serve as a fall-back option when the training data is sparse.

These initial distributions may also be modified depending on the type of speech that will be used with the recognizer. If it is read-speech or instruction-oriented speech produced by standard American-English speakers, then the distributions of stop-like /ð/ may be set higher in all contexts than if the recognizer is to be used for non-task-oriented spontaneous speech of central Ohio residents.

The results of this thesis research can further benefit feature-based ASR systems, whether they are statistically based or not. The next section will explore the applications in such systems in detail.

Feature-Based ASR Systems

Most feature-based systems represent phonetic segments with bundles of features (Hou, Rabiner, Dusan, 2006; Launay et al, 2002; Niyogi, 2004; Stevens, 2002; Tang, Seneff, Zue, 2003). These features are the basic units of acoustic modeling and may be grouped into natural classes, such as place of articulation features and manner of production features. For example, manner of production features may consist of vowel, fricative, affricate, nasal, stop, and sonorant. Place of articulation features may consist of labial, dental, alveolar, retroflex, palatal, and velar.

Past observations have indicated that the fricative manner of production for /ð/ is highly variable in English. For example, /ð/ may become nasalized when preceded by /n/, more sonorant and lateralized when preceded by /l/, and stop-like when preceded by /t/. However, the results of this thesis research, along with past research on modified cases of /ð/, indicate that the dental place of articulation seems almost always preserved despite modification in manner of production. These findings suggest that the surface signal is more reliable in providing information on the underlying place of articulation features of /ð/ than on the underlying manner of production features. Thus, the place features and manner features need to be considered separately in the initial stages of a feature-based ASR system. For example, the extraction of the sonorant feature from the surface signal of a modified /ð/ should not exclude the consideration of the dental feature simply because there is no sonorant-dental consonant in English.

Some feature-based ASR systems may handle the stop-like /ð/ by representing it as a dental stop consonant, a separate entity from the dental fricative /ð/. In that case, there would be four possible places of articulation for an English stop consonant: dental, alveolar, labial, and velar. The salient acoustic and spectral-moment measures that are found to distinguish between a dental and an alveolar stop in this thesis would be useful in extracting the appropriate place feature for stop-like /ð/. Such ASR systems can be designed to ensure that these measures are obtained from the speech signal in order to better capture the acoustic differences between stop-like /ð/ and /d/.

Another way of handling stop-like /ð/ is to treat it as a surface modification where the underling features are still those of a voiced dental fricative. The results of this research, along with those of others, indicate that the place features of /ð/ can be more reliably deduced from the surface signal than the underlying manner features. Therefore, the initial conflict in the place and manner features obtained from the surface signal, in this case for the stop-like /ð/, can be resolved by placing more weight on the place features and adding some uncertainty to the manner features. This might be accomplished by having non-zero probability of other manner features despite surface acoustic information. That way, the stop-like /ð/ segment may have the initial representation of dental consonant features with high confidence, stop-like manner features with high confidence, and fricative manner features with confidence (while all other features have low confidence). From this initial representation, the appropriate underlying feature representation of the stop-like /ð/, a voiced dental fricative, may be obtained when the lexicon, which contains only the dental fricative, is accessed.

The contextual frequency of stop-like /ð/ may also be incorporated into feature-based ASR systems, if distribution data across contexts are employed as part of the recognition system.

As discussed in this chapter, information on the context, frequency, and acoustics of stop-like /ð/ can be applied in a number of areas in ASR systems. For probabilistic systems, the contextual frequency of stop-like /ð/ is useful in building acoustic models that provide appropriate estimations of phone-distributions across contexts, despite the potential sparseness of training data. Salient acoustic information, present in the consonant burst spectrum and in the following vowel, can be analyzed to distinguish stop-like /ð/ from /d/. Such acoustic information can be used to resolve conflicts in word sequences with roughly equal probability predicted from the language model. In feature-based ASR systems, these acoustic measures can be used to extract the features of /ð/ despite modification.

The next chapter will discuss how the detailed-acoustic-analysis approach of this thesis can be applied in studying other types of speech variations.

VIII. APPLICATIONS IN THE STUDY OF SPEECH VARIATION

The need to study variations in speech is one of the main drives behind this thesis. Although this research consists of a case study of the stop-like modification of /ð/, the detailed-acoustic-analysis approach may be generalized and modified to study other types of speech variations. This chapter will first briefly outline the research approach, followed by an example of how the approach may be used to study other speech variations. The advantages and challenges of the approach will also be discussed.

Acoustic-Analysis Centered Approach

This detailed-acoustic-analysis approach involves examining the modified speech segment, as well as the context in which the variation occurs. The goal is to find out when, why, and how a particular modification occurs, especially if any characteristics of the canonical form of the phoneme are preserved despite modification. This approach aims not only to gain an understanding of the particular speech variation in question, but also to provide useful information for technological and clinical applications. Elements of this approach have been used by a number of past studies on the acoustics, perception, and automatic classification of various phonemes and phonetic variations (Cao, 2002; Manuel, 1995; Manuel and Wyrick, 1999; Slifka et al., 2004; Stevens, 2002; Suchato, 2004).

This approach begins with a set of hypotheses on the context, acoustic characteristics, and causes of a particular phonetic variation, based on past observations, studies, and theoretical calculations. These hypotheses may include: 1) the context and frequency of the variation in question; 2) the expected acoustic differences when it is compared with other close-sounding and/or confusable phonemes; 3) whether the acoustic differences indicate the preservation of certain characteristics seen in the canonical form; 4) the robustness of salient acoustic measures, if they exist, in automatically classifying the modified case from other close-sounding and/or confusable phonemes. Through the testing of the hypotheses, additional hypotheses may be developed as the data provide further insights into the nature of the variation.

The following is an example of how this approach can be used to study a different type of speech variation.

Example of Phonetic Modification: /t/ with No Release Burst

Figure 8.1 shows the spectrograms of "can't go" and "can go" taken from speech recorded from a female subject at the MIT RLE Speech Communication Group. The figure shows that /t/ in "can't go" is phonetically modified in that its release burst is absent. At first glance, the acoustic signal contains little evidence of the /t/. Formant movements may indicate an alveolar place of articulation for the segment(s) following the vowel in "can't". However, the presence of /t/ does not seem immediately apparent from the formant movements alone. Despite this modification, human listeners do not seem to have trouble in correctly perceiving "can't" when the /t/ burst is absent. Therefore, the acoustic signal must contain information that suggests the presence of /t/.

The first question that this detailed-acoustic-analysis approach would investigate is: What is the context and frequency of the burst-less /t/? It may be hypothesized that the burst-less /t/ occurs with high frequency, but only in contexts where it is preceded by /n/ within the same word. This hypothesis may be tested through examining the frequency of the burst-less /t/ in various contexts, for different types of speech. Such data may also provide insights as to why that particular modification occurs. From there, further experiments studying the mechanisms of this particular variation can be designed and pursued.

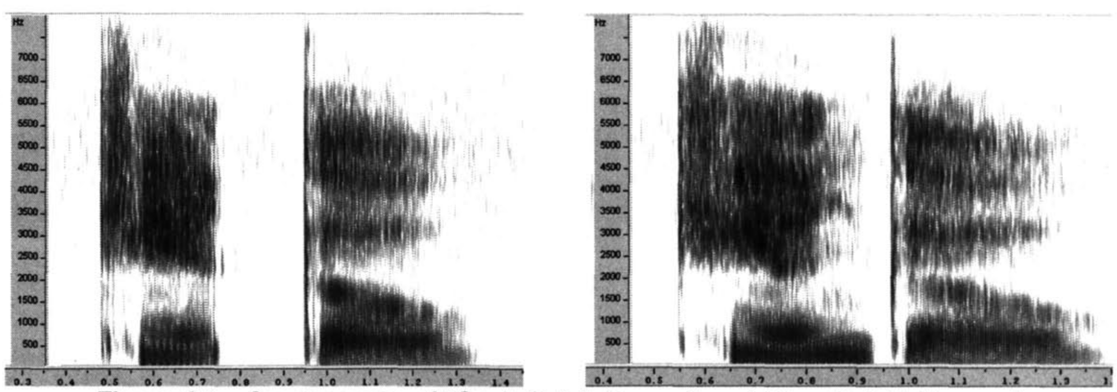


Figure 8.1: Spectrograms of "Can't Go" (left panel) and "Can Go" (right panel) produced by female speaker. The /t/ release is absent in "can't go", however listeners are still able to perceive the utterance correctly.

As previously mentioned, human listeners do not seem to have trouble perceiving "can't" when the /t/ is burst-less. Thus, the next set of questions to be investigated would be: What are the acoustic cues that suggest the presence of /t/? Where can they be found? It may be hypothesized that despite the absent t-burst, acoustic cues that distinguish the sequence from a word-final /n/ can be found in the /n/-to-/t/ transition. It may also be hypothesized that the spectrum taken during that /n/-to-/t/ transition contains information signaling the presence, as well as the place and voicing, of the following consonant segment, in this case the burst-less /t/. The duration of the /n/ segment may indicate whether it is followed by a voiceless consonant. These hypotheses may be tested through the detailed acoustic analysis of the /n/-to-burst-less-/t/ transition in an utterance such as "can't go" with the word-final /n/ in "can go".

If salient acoustic cues that signal the presence of the burst-less /t/ are found, then the next step may be to test their robustness in automatically detecting /t/ despite the absent t-release burst. Another possibility is to perform human perception tests where the salient acoustic cues are varied independently in synthesized tokens. Poor performance results may suggest that additional information exists in the signal. Depending on the results, this approach may further examine where and what the salient acoustic cues are in signaling the presence of /t/ despite the absence of its release burst.

Advantages and Challenges

Since the speech signal is a manifestation of production mechanisms, the non-invasive method of detailed acoustic analysis is an effective first step in uncovering answers to when, why, and how certain phonetic modifications occur in casual speech. The speech recordings used for acoustic analysis are expected to be more natural; subjects speak into a microphone and are most likely not restricted by or attached to some other measuring apparatus. Therefore, this non-invasive approach can be used to study and explore the research problem before conducting more invasive experiments on human subjects. This way, invasive experiments would be performed only out of necessity. Furthermore, such experiments will be designed and performed with prior knowledge obtained from initial acoustic analysis.

The detailed-acoustic-analysis approach may not be solely sufficient in uncovering the exact mechanisms underlying certain speech variations. Speech is produced from an interaction of a multitude of factors. Thus, it might be difficult to isolate the effects of any one factor based only on the acoustic signal. As discussed in the previous chapters, the stop-like /ð/ is one such case where there might be a number of potential factors. It may be necessary to perform physiological measurements in order to isolate and quantify the extent of various factors in causing a particular speech variation.

A major criticism of this detailed-acoustic-analysis approach is that it is time-intensive. From an application point of view, it is much faster to build models of all the phonemes and their common variations, using the same set of parameters; then, those models can be used to detect and handle speech variations. However, modeling without detailed analysis does not provide knowledge on the nature of the speech variation(s). Instead of quick solutions, this approach aims to gather knowledge on speech variations that can be applied in many areas, whether it is for speech recognition, speech transcription, or speech-disorder diagnosis and treatment. The acoustic-detail-oriented approach may take more time to undertake, but the potential rewards are expected to be greater in scope and wider in applicability.

The next chapter will discuss future directions of this thesis research.

IX. FUTURE DIRECTIONS IN THE STUDY OF STOP-LIKE /ð/

One future direction of this thesis work may concentrate on teasing apart the potential causes of the stop-like modification of /ð/. As previously discussed, the context of stop-like /ð/ suggests that the modification may have a number of contributing factors, including physiological mechanisms of speech production and prosody. This chapter will outline some initial steps in investigating these possible contributors to the modification. Other extensions on this thesis, which may lie in areas of acoustical analysis of American-English dialects and the development of acoustic measures for speech disorder diagnosis, will also be discussed

Physiological Effects On Stop-like /ð/

Physiological mechanisms of speech production most likely play a large role in the stop-like modification of /ð/. Such mechanisms involve the complex interaction of articulatory movements, aerodynamic forces, and biological properties of the vocal apparatus. An initial step in investigating physiological factors may be to measure the timing of the articulatory transition from a preceding obstruent into a stop-like /ð/ and a continuant /ð/. The goal is to find whether there is any correlation between the timing of the articulatory transition into /ð/ and the stop-like modification of /ð/.

It may be hypothesized that the articulatory transition time from an obstruent into a stop-like /ð/ is shorter than that into a continuant /ð/. In such sequences, intra-oral pressure is built up during the preceding obstruent's period of closure or constriction. When the preceding consonant is released, intra-oral pressure is also released (Stevens, 1998).

Therefore, the necessary intra-oral pressure for maintaining a /ð/ constriction needs to build up after the preceding consonant's release. Under this hypothesis, the articulators transitioning from the release of a preceding obstruent to a stop-like /ð/ most likely reach the target position before enough intra-oral pressure is built up to maintain a constriction.

The presence of intra-oral pressure plays an important role in the maintenance of a fricative constriction. It applies downward force on the tongue surface and force leading to some displacement of the vocal tract walls. Thus, a constriction area may tend to increase with increased intra-oral pressure (Badin, 1989; Stevens, 1998). In the absence of intra-oral pressure, the static area of the oral constriction may be zero or negative (pushed up further up against the palate). Thus, insufficient intra-oral pressure may lead to the collapse of the constriction into a complete closure during the production of a stop-like /ð/.

Preliminary measurements on the timing of articulatory transition from a preceding obstruent to /ð/ have been made on forty-five k#ð sequences of TIMIT Directory 7. The sequence is taken from "like that" in the sentence "Don't ask me to carry an oily rag like that". Each sequence is produced by a different speaker. Since every TIMIT speaker reads this sentence, there are many more tokens of this particular k#ð sequence than other contexts where /ð/ is preceded by a stop-consonant. A number of the k#ð sequences are not included because the /k/ release is absent.

The articulatory-transition time is measured from the release of the preceding /k/ to the beginning of the following /ð/. As shown in **Figure 9.1**, this is measured from the onset of the /k/-release burst to the beginning of the stop gap for the stop-like /ð/. For the voiced continuant /ð/ shown in **Figure 9.2**, this is measured from the beginning of the /k/-release burst to the beginning of the voiced frication of /ð/. As for the unvoiced continuant /ð/ in **Figure 9.3**, time is measured from the beginning of the k-release burst to the beginning of the /ð/ frication. It is noted that the waveforms may not always display clear-cut information on the beginning and end of speech events. The author did her best to be accurate and consistent in making the measurements.

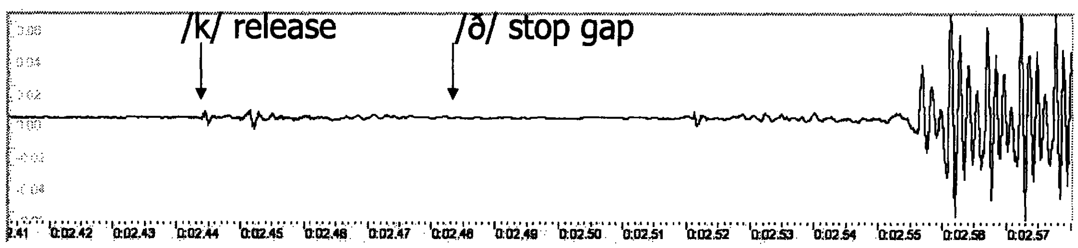


Figure 9.1: Waveform of /k/-release burst to the beginning of the stop gap for a stop-like /ð/.

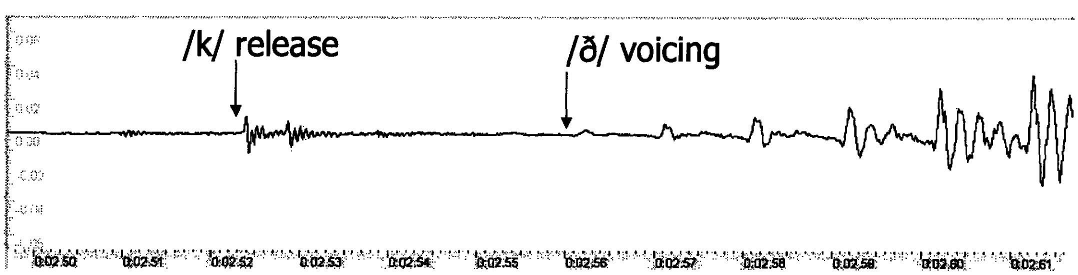


Figure 9.2: Waveform of /k/-release burst to the beginning of frication in a voiced continuant /ð/.

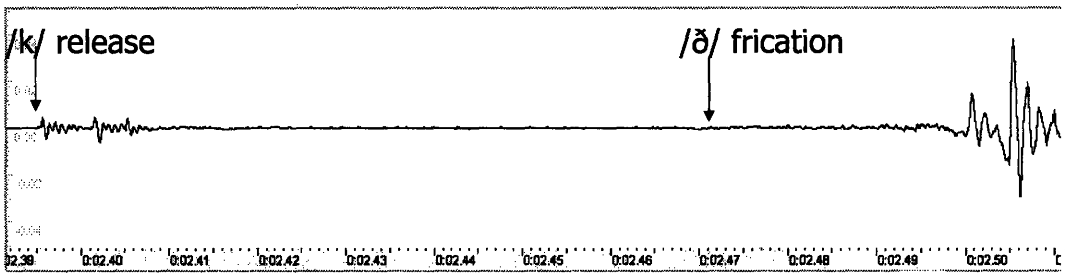


Figure 9.3: Waveform of /k/-release burst to the beginning of frication in an unvoiced continuant /ð/.

Table 9.1 shows the average time of the transition from the /k/ to the three types of /ð/. Of the twenty-five stop-like /ð/ cases analyzed, the time from the /k/ release to the beginning of the stop gap averaged 10ms. The time from the /k/ release to the beginning of voiced frication of five cases of voiced continuant /ð/ averaged 41ms; the time from the /k/ release to the beginning of the frication averaged 40ms for the fifteen unvoiced continuant /ð/ cases.

	Transition Time Duration (Average)	**Transition Time Duration (Min and Max)**
/k/ release to beginning of stop-like /ð/ stop-gap (25)	10ms	4ms to 40ms
/k/ release to beginning of voicing of voiced continuant /ð/ (5)	41ms	19ms to 71ms
/k/ release to beginning of frication of unvoiced continuant /ð/ (15)	40ms	13ms to 83ms

Table 9.1: Average time from /k/ release to stop-like /ð/, voiced continuant /ð/, or unvoiced continuant /ð/, in the k#ð sequence.

The results show that the articulatory transition time from /k/ to the stop-like /ð/ is much shorter than to the continuant /ð/. This supports the hypothesis that the target position for /ð/ may be reached too early, thus not allowing enough time to build up sufficient intra-oral pressure for maintaining a constriction.

However, there is notable overlap in the articulatory timing data among the three groups. This finding suggests that there is no clear time threshold that must be met in order to produce a continuant /ð/. If articulatory timing is an accurate predictor of intra-oral pressure, then the overlap in the timing distributions suggests that other factors may also be influencing stop-like /ð/ production; such factors may include active muscle control, biological properties of the vocal apparatus, or higher level speech planning. Furthermore, the relationship between articulatory timing and stop-like /ð/ can only be used to deduce articulator positions and intra-oral pressure. It does not reveal the exact positions of the articulators or the precise intra-oral pressure during /ð/ production, in particular during the stop gap.

Based on the articulatory timing results, it may be fruitful to conduct physical experiments to measure the intra-oral pressure and the articulator positions during /ð/ production. One way to measure the intra-oral pressure is to place a measuring device behind the point of /ð/ closure in the mouth. Since the point of contact is between the tongue tip and the back of the front teeth, the device would most likely be placed in the subject's mouth via the nose. Then, the device can measure intra-oral pressure while the subject is producing utterances containing /ð/. This procedure is most likely uncomfortable and invasive for human subjects; a medical professional should be present to supervise such an experiment.

The articulator positions during /ð/ production may be obtained through electro-palatography, which determines the tongue-palate contact during speech. This method uses an artificial palate with electrodes embedded on the tongue-facing side. The palate is custom-made for each subject. Contact information is collected by the computer as the subject produces speech while wearing the artificial palate. This method would provide information on the place and time of the palato-lingual contact during stop-like /ð/ production.

Prosodic Effects On Stop-like /ð/

Physiological factors may not be the only ones influencing the production of stop-like /ð/. An upward trend is observed in the prosodically labeled AEMT recordings, where the stop-like occurrence increases with higher positions on the prosodic hierarchy. In addition, speakers are found to produce stop-like /ð/ more often in the read speech of TIMIT and the task-oriented spontaneous speech of AEMT than in the non-task-oriented spontaneous speech of Buckeye. These results suggest that the stop-like modification of /ð/ may be influenced by prosody and/or other aspects of speaking style and manner.

A next step may be to conduct in-depth analysis of the stop-like modification of /ð/ in relation to the prosodic location of /ð/. Although comparisons have been made for /ð/ tokens across prosodic domains in the AEMT database, the segmental contexts of those comparisons are not kept constant. There are not enough AEMT tokens at the small-phrase boundary and mixed-case boundary positions to make such a comparison while keeping the segmental context constant. Thus, data need to be obtained from a larger, prosodically-labeled database; one possible candidate is the Boston University News Radio Corpus (Ostendorf, Price, and Shattuck-Hufnagel, 1995).

If stop-like /ð/ occurrence is found to consistently increase with higher positions on the prosodic hierarchy, then it would be fruitful to ask why that relationship exists. It might be hypothesized that stop-like /ð/ is a form of articulatory strengthening. Articulatory strengthening is typically characterized as greater palato-lingual contact; it is found to occur with more frequency in constituents that are higher in the prosodic hierarchy than those that are lower (Fougeron and Keating, 1997; Keating, Cho, Fougeron and Hsu, 2003). Despite fitting the characteristics of articulatory strengthening, it still unclear as to whether stop-like /ð/ should be considered as a strengthened form of /ð/.

This issue may be examined through human perception tests. If stop-like /ð/ is a strengthened form, then listeners should be more accurate in perceiving /ð/ when it is stop-like than when it is not. This hypothesis may be tested by using contextually-ambiguous phrases such as “I say, the seat is dirty” and “I say, deceit is dirty”; “I say, the feet stinks” and “I say, defeat stinks”; “I say, the Nile runs deep” and “I say, denial runs deep”. There should be two versions for each phrase containing the function word “the”; one version will contain a stop-like /ð/ and the other will contain a continuant /ð/. The phrases that start with /d/ serve as the control group. Subjects will listen to these contextually-ambiguous phrases one at a time and indicate which one of the two possible phrases they heard. If stop-like /ð/ is a strengthened form, then listeners should have higher accuracies in perceiving phrases with a stop-like /ð/ than those with a continuant /ð/. This experiment will also provide data on how accurate humans are in the perception of /d/ versus stop-like /ð/ when the two consonants are in contextually ambiguous phrases. This information may be used as a basis of comparison with automatic classification of stop-like /ð/ and /d/.

Another possible perception experiment to test whether stop-like /ð/ is a strengthened form is to play two versions of the same phrase, one with a stop-like /ð/ and the other with a continuant /ð/. Then, the listener has to indicate which phrase sounds clearer. If stop-like /ð/ is a strengthened form, then listeners should indicate that phrases with a stop-like /ð/ are clearer than those with a continuant /ð/.

Study of American-English Dialects

In addition to the study of possible causes of stop-like /ð/, future extensions of this work may focus on stop-like modification of dental fricatives or dentalization of alveolar consonants in other dialects of American-English. For example, speakers of African-American English typically replace word-initial /ð/ with [d] and word-medial /ð/ with [d] or [v] (Green, 2002). Based on the results of this research, if the [d] substitute for /ð/ in African-American English has a dental place of articulation, then it should be significantly different from the alveolar /d/ in burst spectrum shape, burst peak frequency, and second formant at following-vowel onset. These spectral measures can also be used to study stop-like modification of dental fricatives and the dentalization of alveolar stop consonants in the New York dialect of American English (Labov, 1982). Such studies would not only add to the understanding of American-English dialects, but could also be useful in the development of ASR systems that are robust in recognizing different dialects of speech.

Clinical Applications

Additional extensions of this research may focus on the development of acoustic measures for speech disorder diagnosis. For example, if a child consistently produces stop-like /ð/ instead of continuant /ð/, it may be assumed that he is replacing /ð/ with /d/. Before jumping to that assumption, it would be important to know the place of articulation for the stop-like consonant that the child is producing. The acoustic measures found in this study can be used to compare the child's stop-like /ð/ with his/her /d/. The acoustic differences between the two sounds, if they are found, should provide information on the child's place of articulation for his/her stop-like /ð/. That way, the clinician can determine whether the child has trouble with the place of articulation as well as the manner of production for /ð/. This information may be essential in the correct diagnosis and eventual treatment of a speech problem.

X. SUMMARY

This thesis has focused on the stop-like /ð/, a common variation of a frequently occurring sound in spoken English. The variation exhibits a drastic change from the canonical /ð/; the manner of production is changed from one that is fricative to one that is stop-like. Furthermore, the place of articulation of stop-like /ð/ has been a point of uncertainty, leading to the confusion between stop-like /ð/ and /d/. In order to better understand the nature of this common variation, the three major studies of this thesis have been designed to uncover the context and place of articulation of stop-like /ð/, as well as the salient acoustic cues that distinguish between stop-like /ð/ and /d/. The immediate application of the thesis findings lies in automatic speech recognition, speech transcription, and clinical diagnosis and treatment of speech disorders. In addition, the detailed-acoustic-analysis approach of this thesis may be generalized and modified to study other types of speech variations. The following is a summary of the thesis findings, the potential application of the research results and research approach, as well as future extensions on this thesis work.

Contextual Frequency of Stop-like /ð/

The first study of this thesis has focused on the contextual frequency of stop-like /ð/ across different speech databases. Word-initial /ð/ in the read speech of TIMIT, the task-oriented spontaneous speech of AEMT, and the non-task-oriented spontaneous speech of Buckeye are examined acoustically. Word-initial /ð/ is categorized as stop-like when there is acoustic evidence to indicate that a complete closure is made and released. This evidence involves a release burst that is preceded by a stop gap or a decreasing voice bar in the acoustic signal.

Stop-like /ð/ is found to occur, to varying degrees, in all types of contexts across the three databases. The stop-like modification occurs most often when /ð/ is preceded by silence (77% in TIMIT, 79% in AEMT, and 45% in Buckeye) or when it is preceded by a stop consonant (73% in TIMIT, 88% in AEMT, and 33% in Buckeye). Stop-like /ð/ occurrence is less frequent (28% in TIMIT, 33% in AEMT, and 9% in Buckeye) when the consonant is preceded by a fricative or an affricate. The stop-like modification rarely occurs (14% in TIMIT, 16% in AEMT, and 7% in Buckeye) when /ð/ is preceded by a vowel or liquid consonant.

Comparing the three databases, speakers produced stop-like /ð/ much more often in read speech (TIMIT) and task-oriented spontaneous speech (AEMT) than in non-task-oriented spontaneous speech (Buckeye). This finding does not support the hypothesis that stop-like /ð/ occurrence increases with the level of casualness in speech. Furthermore, an upward trend is observed in the prosodically-labeled AEMT recordings, where the stop-like occurrence increases with higher positions on the prosodic hierarchy.

These findings suggest that the stop-like modification may have a number of contributing factors, including physiological mechanisms of speech production, prosody, dialect, and/or other aspects of speaking style and manner.

Acoustics of Stop-like /ð/

The second study of the thesis involves the acoustic analysis of stop-like /ð/, in comparison with /d/. Acoustic measures are obtained in the burst spectrum and at the following vowel onset for 100 word-initial stop-like /ð/ tokens and for 102 /d/ tokens, all from TIMIT. All cases are followed by an unstressed vowel, [ə] or [ɨ], and extracted from utterances of 23 female and 59 male speakers. Each speaker has at least one /ð/ case and one /d/ case included in this analysis.

On average, stop-like /ð/ is found to have higher burst spectrum peak (F_{hi}) and higher mean frequency in the burst spectrum above 2kHz than /d/, both with statistical significance. The differences would be expected if the dental place were preserved for stop-like /ð/; dentals have higher front cavity resonances because of their shorter front cavity than that of alveolars. In addition, stop-like /ð/ is found to have lower F2 at vowel onset than /d/, with statistical significance. This difference is also expected if the dental place were preserved for the modified /ð/; dentals have a lower and more backed tongue body, as well as a slightly longer back cavity, than alveolars.

Statistically significant differences are also observed in the spectral amplitudes and shapes of the two consonants. On average, stop-like /ð/ has smaller ratio of peak amplitude above 2kHz to that below 2kHz (A_{hi}-to-A_{lo}), smaller ratio of total amplitude above 2kHz in the consonant to that in the vowel (C-to-V), larger standard deviation around the mean frequency, smaller coefficient of kurtosis and smaller coefficient of skewness than /d/. These differences are expected if stop-like /ð/ has a dental place of articulation. Dental consonants have weak energies and flatter spectral shape at high frequencies; whereas alveolar consonants have greater energies and more compact shape at high frequencies.

Overall, the analysis results support the hypothesis that the dental place of articulation is preserved for stop-like /ð/. In addition, data for each acoustic and spectral measure demonstrate statistically-significant differences between stop-like /ð/ and /d/. Thus, all measures seem salient in distinguishing stop-like /ð/ from /d/.

Automatic Classification of Stop-like /ð/ and /d/

The third study uses these salient measures, taken only from the consonant-burst and at the following vowel onset, in the automatic classification of stop-like /ð/, distinguishing it from /d/. Automatic classification is performed on 100 stop-like /ð/ tokens and 102 /d/ tokens with similar following-vowel context. These cases are obtained from utterances of 23 female and 59 male speakers in the TIMIT database. Each speaker has at least one /ð/ case and one /d/ case included in this study.

Two different classification algorithms are employed. The first algorithm involves leave-one-out classification using the PCA-transformed data. The feature space consists of the eight principal components obtained from the eight salient measures analyzed in the previous study of this thesis; they are F_{hi}, A_{hi}-to-A_{lo}, C-to-V, F2 at following vowel onset, mean frequency in burst spectrum above 2kHz, standard deviation around that mean frequency, coefficient of skewness, and coefficient of kurtosis. The classification decision is based on the Euclidean distance of the unknown token from the centroids of stop-like /ð/ and /d/ data. This algorithm yields 83% accuracy in classifying the 100 stop-like /ð/ cases and 86% accuracy in classifying the 102 /d/ cases.

The second algorithm involves leave-one-out classification using LDA; this method yields 88% accuracy in classifying the 100 stop-like /ð/ cases and 80% accuracy in classifying the 102 /d/ cases. The feature space is also obtained from F_{hi}, A_{hi}-to-A_{lo}, C-to-V, F2 at following vowel onset, mean frequency in burst spectrum above 2kHz, standard deviation around that mean frequency, coefficient of skewness, and coefficient of kurtosis.

The 80+% classification accuracies obtained in this study suggest that the eight acoustic cues and spectral moment measures obtained from the consonant segment and its following vowel are robust in automatically classifying stop-like /ð/ and /d/. Therefore, these measures can be used by automatic speech recognizers to correctly identify the stop-like /ð/ despite a change in manner of production.

It is found that factors influencing the first five principal components (which account for roughly 95% of the variance in the data) can be generalized into various classes of acoustic and spectral measures. The first principal component is dominated by the front cavity resonance and correlated measures in the burst spectrum. The second and fifth principal components are dominated by acoustic cues from the burst spectrum. The back cavity resonance is the major contributor to the third principal component. The fourth component is mostly influenced by amplitude measures from the burst spectrum.

These relationships can be used by speech recognizers to group acoustic and spectral measures into larger classes. These classes may be ranked according to their relative contributions to the variance in the data. Doing so would allow information from one or more classes to resolve conflicts or lack of information in another class(es), regardless of the number of measures in each class.

Applications in ASR Systems

Information on the contextual frequency and acoustics of stop-like /ð/ can be applied in a number of areas in ASR systems. For probabilistic systems, the contextual frequency of stop-like /ð/ is useful in building acoustic models that provide appropriate estimations of phone-distributions across contexts, despite the potential sparseness of training data. Salient acoustic information, present in the consonant burst spectrum and at the onset of its following vowel, can be analyzed to distinguish stop-like /ð/ from /d/. Such acoustic information can be used to resolve conflicts in word sequences with roughly equal probability predicted from the language model.

Despite the high variability of /ð/, the results of this thesis research, along with those of other studies, suggest that the dental place of articulation seems always to be preserved despite modification in manner of production. Thus, the surface signal is more reliable in providing information on the underlying place of articulation features of /ð/ than on the underlying manner of production features. Furthermore, the variability of /ð/ indicates that the place features and manner features must be considered separately in a feature-based ASR system. For example, the extraction of the sonorant feature from the surface signal of a modified /ð/ should not exclude the consideration of the dental feature simply because there is no sonorant-dental consonant in English.

It is hoped that the results of this thesis may be incorporated into ASR systems to improve the recognition of one of the most common sounds in English.

Application of Approach in Studying Speech Variations

The need to study variations in speech is one of the main drives behind this thesis research. Although this research consists of a case study on the stop-like modification of /ð/, the detailed-acoustic-analysis approach may be generalized and modified to study other types of speech variations. The approach starts with a set of hypotheses on the behavior, acoustics characteristics, and causes of a particular phonetic variation, based on past observations, studies, and theoretical calculations. Through the testing of the hypotheses, additional hypotheses may be developed as the data provide further insights into the nature of the variation.

Future Directions

Future extensions of this thesis work may concentrate on teasing apart the potential causes of stop-like /ð/. The findings of this thesis suggest that the modification may have a number of contributing factors, including physiological mechanisms of speech production, prosody, and other aspects of speaking style and manner. Acoustic analysis, along with physical measurements of intra-oral pressure and articulatory movements, may be conducted to deduce the role and extent of aerodynamic, articulatory and other physiological effects on the production of stop-like /ð/.

Prosodic effects may be studied through the comparisons of stop-like /ð/ frequency across prosodic domains, with the same segmental context. This analysis needs to be conducted using a large prosodically labeled database to ensure sufficient numbers of tokens for each domain while the segmental context remains the same. Perception tests may also be administers to test whether stop-like /ð/ is a form of articulatory strengthening.

It is hoped that this thesis research will be one among many studies that focus on uncovering the mysteries behind common phonetic variations in everyday speech.

XI. REFERENCES

Badin, P. (1989). Acoustics of voiceless fricatives: Production theory and data. *Speech Transmission Laboratory Quarterly Progress and Status Report 3,* Royal Institute of Technology, Stockholm, 33-55.

Balakrishnama, S., Ganapathiraju, A. (1998). *Linear Discriminant Analysis - A Brief Tutorial*, Institute for Signal and Information Processing.

Bertsekas, D. P., Tsitsiklis, J.N. (2002). *Introduction to Probability*. Belmont, MA: Athena Scientific.

Bybee, J. (2001). *Phonology and language use*. Cambridge University Press.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Raymond, W. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of Acoustical Society of America*, **113**(2), 1001-1024.

Cao, A.Y. (2002). Analysis of acoustic cues for identifying the consonant /ð/ in continuous speech. Master's dissertation, Massachusetts Institute of Technology.

Campbell, J.P., Reynolds, D.A. (1999). Corpora for the evaluation of speech recognition systems, IEEE International Conference on Acoustics, Speech, and Signal Processing, Phoenix, AZ.

Cho, T., Keating, P. (2001). Articulatory and acoustic studies on domain-initial strengthening in Korean, *Journal of Phonetics,* **29**, 155-190.

Degenshein, R. (2004). Dholuo interdentals: fricatives or affricates? Evidence from domain-initial strengthening. *Conference Proceedings of From Sound to Sense*, Massachusetts Institute of Technology, Cambridge, MA.

Denes, P. B. (1963). On the statistics of spoken English. *Journal of Acoustical Society of America*, **88**(6), 894 and 897.

du Preez, J.(2005). Pattern Recognition. 414/813 Course Lecture Notes, University of Stellenbosch.

Fant, G. (1960) *Acoustic Theory of Speech Production*. Netherlands: Mouton & Co.

Fisher, W., Zue,V., Bernstein, J., Pallet, D. (1987). An acoustic-phonetic database. *Journal of Acoustical Society of America*, **81**, S92-S93.

Forrest, K., Weismer, G., Milenkovic, P., and Dougall, R. M. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of Acoustical Society of America*, **84**, 115-123.

Fougeron, C. and Keating, P. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of Acoustical Society of America*, **101(6)**, 3728-3740.

Gimson, A. C. (1989). *An introduction to the pronunciation of English.* London: Edward Arnold.

Glass, J.R. (1988). Finding acoustic regularities in speech: applications to phonetic recognition. PhD dissertation, Massachusetts Institute of Technology, Cambridge MA.

Glass, J., Zue, V. (2005). Automatic Speech Recognition. 6.345 Course Lecture and Notes, Massachusetts Institute of Technology.

Green, L. J. (2002) *African American English: A Linguistic Introduction.* Cambridge: Cambridge University Press.

Hou, J., Rabiner, L., Dusan, S. (2006). Automatic Speech Attribute Transcription (ASAT)-the front end processor. IEEE International Conference on Acoustics, Speech, and Signal Processing, Toulouse, France.

Huang, X., Acero, A., Hon, H.W. (2001). *Spoken language processing*. Prentice Hall.

iSixSigma Staff. (2003, May 6). Sample Size... Why 30?. *iSixSigma.com.* Retrieved May 29, 2007, from http://www.isixsigma.com/library/content/c030506a.asp

Jolliffe, I.T. (1986). *Principal component analysis*, Springer, New York.

Jongman, A., Blumstein, S. E., Lahiri, A. (1985). Acoustic properties for dental and alveolar stop consonants: a cross-language study. *Journal of Phonetics*, **13**, 235-251.

Jongman, A., Wayland, R., Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of Acoustical Society of America*, **108**(3), 1252-1263.

Jurafsky, D., Bell, A., Fosler-Lussier, E., Girand, C., Raymond, W. (1998). Reduction of English function words in switchboard. *Proceedings of the International Conference of Spoken Language Processing,* Sidney, Australia.

Keating, P., Cho, T., Fougeron, C., and Hsu, C. (2003). Domain-initial articulatory strengthening in four languages. *Phonetic Interpretation: Papers in Laboratory Phonology 6*, edited J. Local, R. Ogden, R. Temple, Cambridge University Press, pp.143-161.

Labov, W. (1982) *The social stratification of English in New York City* Center for Applied Linguistics.

Launay, B., Siohan, O., Surendran, A., and Lee, C.H. (2002). Towards Knowledge-Based Features for HMM Based Large Vocabulary Automatic Speech Recognition. *In Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 817–820.

Lee, K. (1990). Context-dependent phonetic hidden markov models for speaker-independent continuous speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 38, No. 4.

Maddieson, I. (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.

Manuel, S.Y. (1995). Speakers nasalize /ð/ after /n/, but listeners still hear /ð/. *Journal of Phonetics*, **23**, 453-476.

Manuel, S.Y., Wyrick, G.C. (1999). Casual speech: a rich source of intriguing puzzles. *The 14th International Congress of Phonetic Sciences Conference Proceedings*, **1**, 679-682.

McAllister, J., Sotillo, C., Bard E.G., and Anderson, A.H. (1990). "Using the map task to investigate variability in speech," *Occasional paper,* Department of Linguistics, University of Edinburgh.

Morgan, N., et al (2005). Pushing the envelope-aside. *IEEE Signal Processing Magazine, September 2005, 81-88.*

Niyogi, P. (2004). Towards a computational model of human speech perception. *Conference Proceedings of From Sound to Sense*, Massachusetts Institute of Technology, Cambridge, MA.

Ostendorf, M., Price, P. and Shattuck-Hufnagel, S. The Boston University Radio News Corpus. *Boston University Technical Report,* ECS-95-001, 1995.

Paradis, C., Prunet, J.F. (1991). *Phonetics and phonology: the special status of coronals.* New York: Academic Press.

Pastel, L. (1987). *Turbulent noise sources in vocal tract models.* Unpublished SM thesis, Massachusetts Institute of Technology, Cambridge MA.

Phillips, M. (2006). Current state of automatic speech recognition technologies. Lexical Access Meeting Presentation, Speech Communications Group at the Research Lab of Electronics, Massachusetts Institute of Technology, Cambridge MA.

Pitt, M., Johnson, K., Hume, E., Kiesling, S., Raymond, W. (2005). The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication, 45, 89-95.*

Roach, P. (2000). *English phonetics and phonology: A practical course*. Third edition. Cambridge University Press.

Sanders, J. R. (1998). Drawing a sample: how big should my sample be? *The Evaluation Center WMU Bulletin 4*. Western Michigan University. Retrieved on May 29, 2007, from http://www.wmich.edu/nonprofit/Bulletin/bulletin4.htm

Shadle, C. (1985). The acoustics of fricative consonants. PhD dissertation, Massachusetts Institute of Technology, Cambridge MA.

Shadle, C.H., Mair, S.J. (1996). Quantifying spectral characteristics of fricatives. *Proceedings of the International Conference of Spoken Language Processing,* Philadelphia, PA.

Shattuck-Hufnagel, S.,Turk, A.E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research,* vol. 25(2), 193-246.

Shattuck-Hufnagel, S., Veilleux, N. (2000): The special phonological characteristics of monosyllabic function words in English, *Proceedings of the International Conference on Spoken Language Processing,* vol.1, 540-543.

Shockey, L. (2003). *Sound patterns of spoken English.* Blackwell Publishing.

Slifka, J. (2003). Respiratory constraints on speech production: Starting an utterance. *Journal of Acoustical Society of America,* **114**(6), 3343-3353.

Slifka, J., Stevens, K. N., Manuel, S. & Shattuck-Hufnagel, S. (2004). A landmark-based model of speech perception: History and recent developments. *Conference Proceedings of From Sound to Sense*, Massachusetts Institute of Technology, Cambridge, MA.

Stevens, K.N. (1998). *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Stevens, K.N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of Acoustical Society of America*, **111**(4), 1872-1891.

Stevens, K.N., Keyser, S.J.(1989) Primary features and their enhancement in consonants. *Language*, **65,** 1, 81-106.

Stoel-Gammon, C., Williams, K., and Buder, E. (1994). Cross-language differences in phonological acquisition: Swedish and American /t/. *Phonetica,* **51**, 146-158.

Suchato, A. (2004). Classification of stop consonant place of articulation. PhD dissertation, Massachusetts Institute of Technology, Cambridge MA.

Sundara, M. (2005). Acoustic-phonetics of coronal stops: A cross-language study of Canadian English and Canadian French. *Journal of Acoustical Society of America*, **118**(2), 1026-1037.

Tang, M., Seneff, S., and Zue, V., (2003). Modeling Linguistic Features in Speech Recognition,"*Proc. Eurospeech*, 2585-2588, Geneva, Switzerland, September 2003.

Wilde, L. (1995) Analysis and synthesis of fricative consonants. Ph.D. dissertation, Massachusetts institute of Technology.

Young, S. (1996) Large vocabulary continuous speech recognition: A review. *In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding,* pp. 3-28.

Young, S. (2001) Statistical modeling in continuous speech recognition. *Proc. Int Conference on Uncertainty in Artificial Intelligence*, Seattle.

Appendix A: Tube Approximation of /ð/

The following is a tube approximation of /ð/. The configuration can be broken down into simple tube in order to approximate the resonant frequencies.

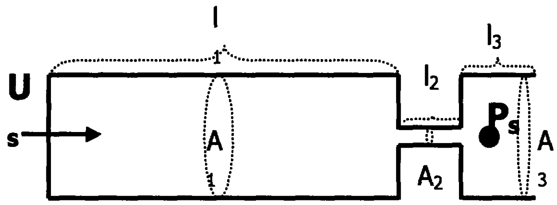
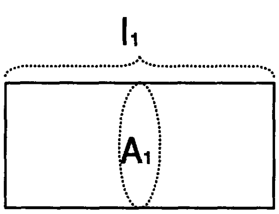


Table A gives the approximate values of the configuration that can be used to calculate the resonant frequencies.

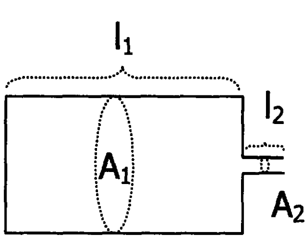
l1	15.8cm		A1	3cm2
l2	0.7cm		A2	0.15cm2
l3	1.2cm		A3	3cm2

Table A. Approximate values deduced from Stevens 1989 & Wilde 1995

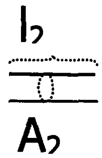
The following are the simple tubes that make up the configuration for /ð/. The equations for calculating each tube's resonant frequencies are shown below.



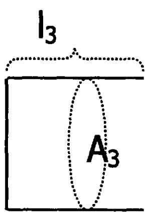
$$f_n = \frac{c}{2l_1} n = 0Hz, 1120Hz, 2240Hz, ...$$



$$f = \frac{c}{2\pi\sqrt{\frac{A_1 l_1 l_2}{A_2}}} = 370Hz$$

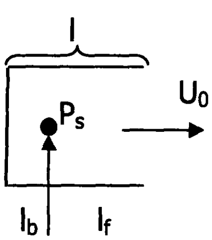


$$f_n = \frac{c}{2l_2} n = 0Hz, 25285Hz, \ldots$$

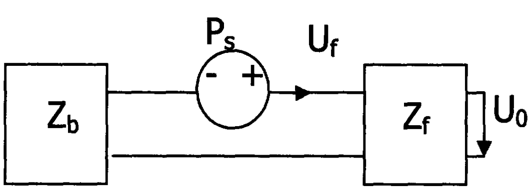


$$f_n = \frac{c}{4l_3}(2n-1) = 7375Hz, \ldots$$

The volume velocity U, representing the glottal source, excites these resonant frequencies of the configuration. However, the pressure source P_s, representing the noise source, excites mainly the front cavity. The derivation of the excited frequencies, f_n, and zeros of the front cavity, z_n, is below. Zeros are frequencies where there is no energy output.



$$Z_b = -j\frac{\rho c}{A}\cot kl_b \qquad Z_f = j\frac{\rho c}{A}\tan kl_f$$

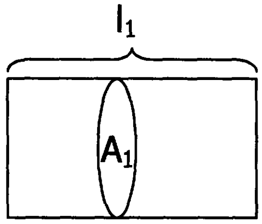


$$U_f = \frac{P_s}{Z_b + Z_f} = j\frac{A}{\rho c}\frac{\sin kl_b \cos kl_f}{\cos k(l_b + l_f)} P_s$$

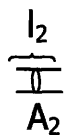
$$\frac{U_0}{U_f} = \frac{U_0}{U_0 \cos kl_f} = \frac{1}{\cos kl_f} \qquad f_n = \frac{c}{4l}(2n-1)$$

$$\frac{U_0}{P_s} = \frac{U_0}{U_f} \cdot \frac{U_f}{P_s} = j\frac{A}{\rho c}\frac{\sin kl_b}{\cos kl} \qquad z_n = \frac{c}{2l_b} n$$

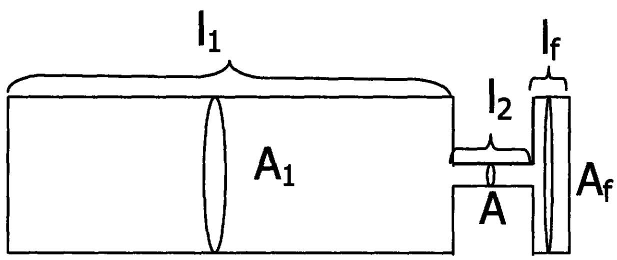
The presence of P_s also leads to additional zeros of the configuration. They are shown below.



$$z_n = \frac{c}{2l_1} n$$



$$z_n = \frac{c}{2l_2} n$$



$$Z = \frac{1}{j\omega C_{total}} + j\omega M_c = 0$$

$$\frac{1}{C_{total}} = \frac{1}{C_b} + \frac{1}{C_f} = \frac{C_b + C_f}{C_b C_f} \approx \frac{1}{C_f}$$

$$z = \frac{c}{2\pi}\sqrt{\frac{1}{C_{total} M_c}} = \frac{c}{2\pi}\sqrt{\frac{A_2}{A_f l_2 l_f}}$$

It is noted that the poles and zeros of the back cavity more or less (although not completely) cancel.

Appendix B: Excitation of Resonators

An acoustic resonator is characterized by its natural frequencies. For a uniform tube shown in Figure 1a, these natural frequencies are where there would be infinite output due to an excitation when loss is not considered. There are distributions of sound pressure and volume velocity along the length of the resonator that correspond to each natural frequency; they are called normal modes. The distributions display properties of standing waves. At a particular point along the length of the tube, pressure and volume velocity distributions are 90-degrees out of phase. As illustrated in Figure 1b, the volume velocity amplitude is a maximum at the open end where the pressure amplitude is zero. At the closed end, the pressure amplitude is a maximum while the volume velocity amplitude is zero. The equations* for pressure and volume velocity distributions can be derived from one-dimensional wave propagation through simple tubes, since the cross-sectional area of the tube is small compared to the wavelength.

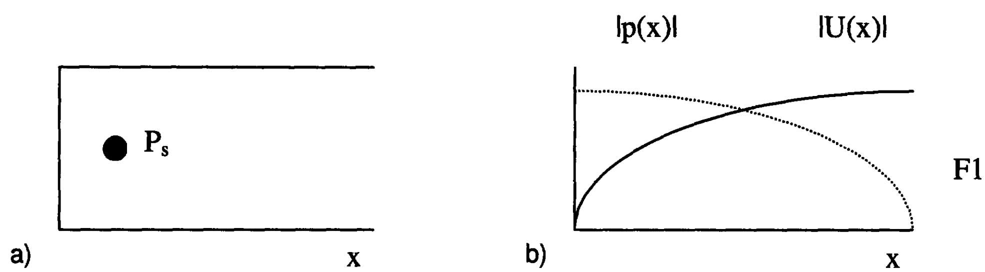


Figure B1. a) Uniform tube with a pressure source P_s. This may represent the front cavity with a noise source in fricative production. b) Distribution of sound pressure amplitude |p(x)| in dotted line and volume velocity amplitude |U(x)| in solid line for the first natural frequency in the uniform tube.

The general principles governing the excitation of normal modes state that an excitation source placed in the resonator will excite the normal modes differently depending on the location of the source (Stevens, 1998). A pressure source will excite a normal mode in proportion to the relative amplitude of the volume velocity distribution at the point where the source is placed. A volume velocity source will excite a normal mode in proportion to the relative amplitude of the pressure distribution at the same location.

The intuition behind the principle is that a pressure source at a location where there is higher volume velocity amplitude for a normal mode will see lower acoustic impedance. Lower acoustic impedance leads to greater volume velocity output at the particular natural frequency. For a tube shown in Figure 1, the farther the pressure source from the closed end, the greater the volume velocity amplitude for F1 and the lower the acoustic impedance; this leads to stronger the excitation of (greater volume velocity output at) the first natural frequency. (The first natural frequency is the main resonance of the front cavity due to the noise source in fricative production.)

* $p(x) = P_m \sin \frac{2\pi f x}{c}; \quad U(x) = jP_m \frac{A}{\rho c} \cos \frac{2\pi f x}{c}$

Appendix C: Radiation Characteristic and Impedance

The radiation impedance due to the mouth opening consists of a resistance and a reactance. It is the load that the surrounding air places on the radiator(s). The resistive part is the radiated energy loss while the reactive term represents the effective mass of vibrating air at the lips. The radiation impedance Z_r may be approximated by that of a circular piston in a sphere. The circular piston represents the mouth while the sphere represents the head. The following is Z_r based on that approximation (Fant, 1960; Stevens, 1998).

$$\frac{P_l}{U_0} = Z_r = \frac{\rho c}{A_m}\left(\frac{\pi f^2}{c^2}A_m\right)K_s(f) + j2\pi f\frac{\rho(0.8a)}{A_m}$$

A_m is the area of the mouth opening, a is the effective radius of the mouth opening, K_s is the head transfer function. P_l is the sound pressure at the lips; U_0 is the volume velocity at the lips.

The radiation resistance is proportional to f^2; this means that radiation loss increases with frequency. The radiation reactance is the energy stored in the acoustic mass; it effectively lengthens the vocal tract and decreases all formants.

Radiation Characteristic

The radiation characteristic represents the relationship between the volume velocity at the lips and the sound pressure at distance r from the lips. At low frequencies (below 4000Hz), the mouth opening can be regarded as a simple source of strength U_0 radiating uniformly in all directions. Then, the sound pressure measured at distance r is:

$$p_r(f) = \frac{j2\pi f\rho}{4\pi r}U_0(f)e^{-j(\frac{2\pi fr}{c})}$$

The radiation characteristic, $R(f)$, is $p_r(f)$ over $U_0(f)$. The magnitude of the radiation characteristic rises with frequency.

$$R(f) = \frac{p_r(f)}{U_0(f)} \qquad\qquad | R(f) | = \frac{f\rho}{2r}$$

Relationship of Radiation Impedance and Radiation Characteristic

The radiation characteristic is the same in the far field for any simple source. However, the radiation impedance differs depending on the type of the source. The radiation impedance indicates the energy radiated from a source to the environment, as well as the resistive and reactive forces of the medium on the source. The resistive forces on the mouth opening increases with frequency; therefore, the energy lost through radiation increases with frequency. This does not conflict with the radiation characteristic. The magnitude of the radiation characteristic rises with frequency because more energy is radiated as frequency increases.

Appendix D: Waveforms of Stop-like and Non-stop-like /ð/

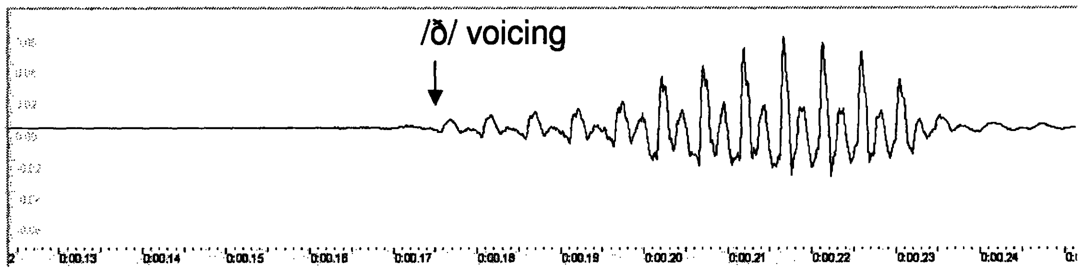


Figure D1. Continuant /ð/ that is preceded by silence, in the utterance-initial "The" (TIMIT fcjs0 sx347.wav).

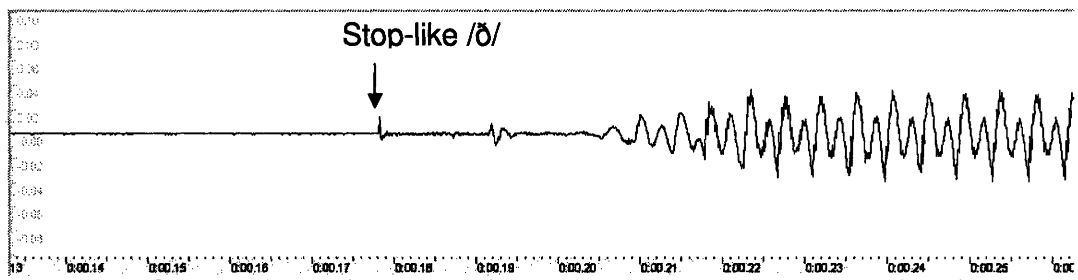


Figure D2. Stop-like /ð/ that is preceded by silence, in the utterance-initial "The" (TIMIT fjen0 sx237.wav). This particular token seems to have two consecutive bursts for /ð/.

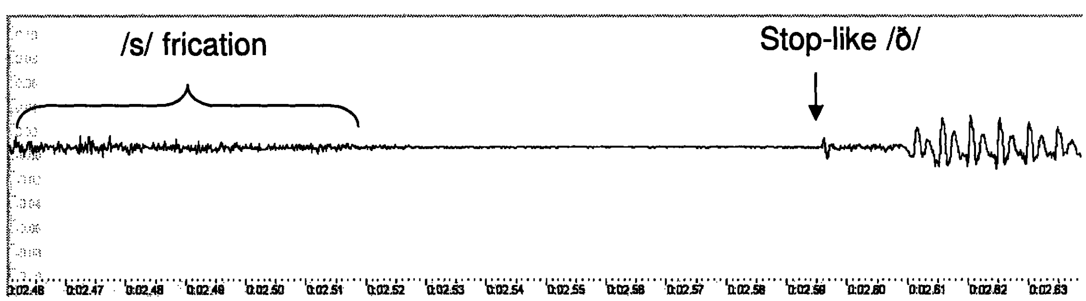


Figure D3. Stop-like /ð/ that is preceded by a fricative /s/ (TIMIT fjen0 sx417.wav).

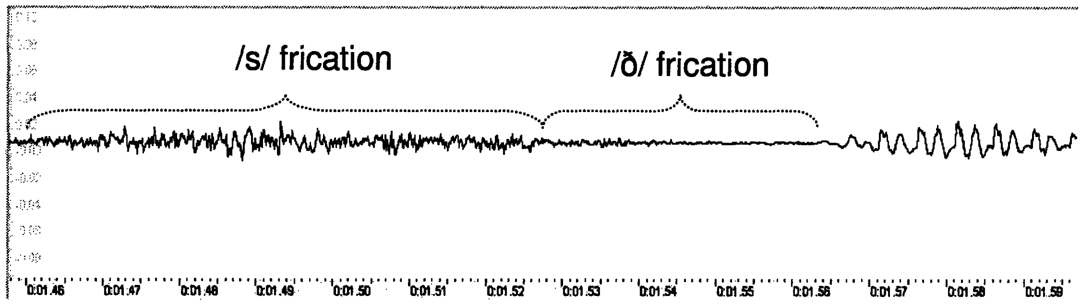


Figure D4. Non-stop-like /ð/ that is preceded by a fricative /s/ (TIMIT fjhk0 sx392.wav). The weak frication of /ð/ immediately follows the /s/ frication. There seems to be no discontinuity in output from the /s/ to /ð/. It is noted that the transition from /s/ to /ð/ may not occur at a discrete point in time. Thus, the brackets for /s/ and /ð/ are dotted.

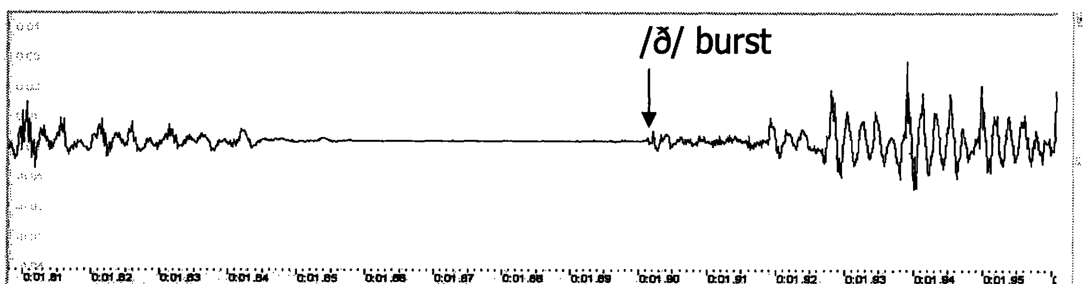


Figure D5. Stop-like /ð/ in "like that" (TIMIT mbsb0 sa2.wav). The preceding stop-consonant has no release burst.

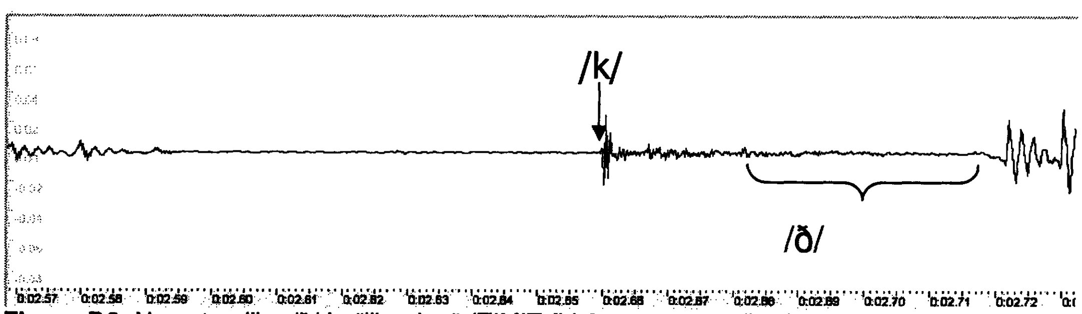


Figure D6. Non-stop-like /ð/ in "like that" (TIMIT fjrb0 sa2.wav). The frication of /ð/ immediately follows the /k/ release. There seems to be no discontinuity in output from the /k/ release to /ð/.

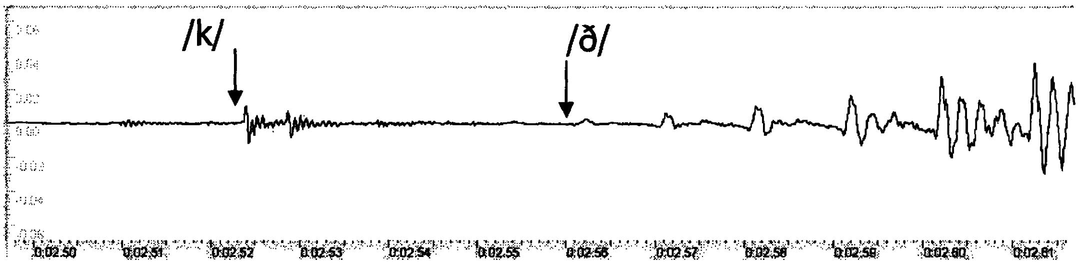


Figure D7. Voiced continuant /ð/ in "like that" (TIMIT mbar0 sa2.wav). The voiced frication of /ð/ is preceded by a period no acoustic output.

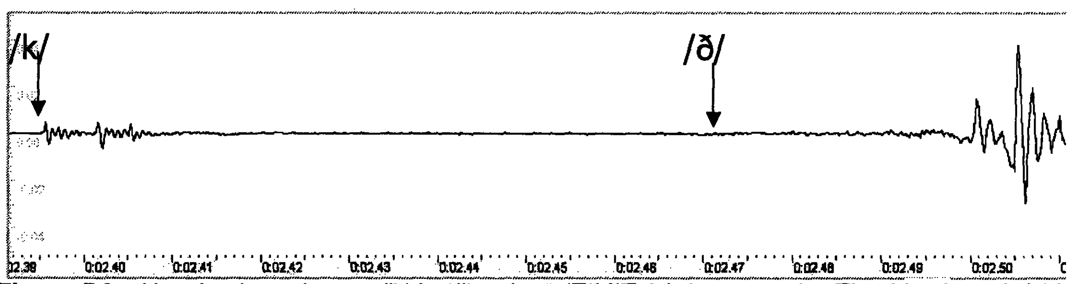


Figure D8. Unvoiced continuant /ð/ in "like that" (TIMIT fcjs0 sa2.wav). The frication of /ð/ is preceded by a period of no acoustic output.

Appendix E: Correlation Coefficient

A correlation coefficient is a number between -1 and 1 which measures the degree to which two variables are linearly related (Bertsekas and Tsitsiklis, 2002). If there is perfect linear relationship with positive slope between the two variables, the correlation coefficient would be 1; if there is positive correlation, whenever one variable has a high (low) value, so does the other. If there is a perfect linear relationship with negative slope between the two variables, the correlation coefficient would be -1; if there is negative correlation, whenever one variable has a high (low) value, the other has a low (high) value. A correlation coefficient of 0 indicates that there is no linear relationship between the variables.

The following equations are used to calculate the correlation coefficient.

$$\rho_{xy} = \frac{Cov(X,Y)}{(\sigma_x \sigma_y)} \text{ where } -1 \leq \rho_{xy} \leq 1$$

$$\text{and } Cov(X,Y) = \frac{1}{n}\sum_{j=1}^{n}(x_j - \mu_x)(y_j - \mu_y)$$

Appendix F: Principal Component Analysis

Principal Component Analysis (PCA) decorrelates the variables of a set of data by projecting the data onto the eigen vectors of its own covariance matrix (du Preez, 2005; Glass and Zue, 2005; Jolliffe, 1986). In addition, PCA chooses a new coordinate system for the data set such that the greatest variance by any projection of the data set comes to lie on the first axis, called the first principal component, the second greatest variance on the second axis, and so on.

The usefulness of PCA are:

- Extraction of independent components from the correlated variables of a data set
- Reduction of dimensionality
- Revelation of relationships among variables

The algorithm of PCA consists of the following steps:

- Calculate covariance matrix Σ for data matrix ***X***
- Do eigen-analysis of $\Sigma = U \Lambda U^T$
- Select M biggest eigen values $\lambda_1 \ldots \lambda_M$
- Pack associated eigen vectors into transformation matrix $W=[u_1 \ldots u_M]$
- Transform data via $\mathbf{Y}=W^T\mathbf{X}$

The drawbacks and limitations of PCA are:

- Dependent only on variance
- Outliers may distort results
- Unless normalized, variables with large variances will dominate the first few principle components
- May miss nonlinear structures
- Does not maximize the between-group separation or minimize with-in group variability

Appendix G: Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) finds a linear subspace that maximizes between-class separability among the feature vector projections in that space (Balakrishnama and Ganapathiraju, 1998; du Preez, 2005). Variables are decorrelated through linear transformations. LDA explicitly aims to model the differences between classes, whereas PCA does not.

The algorithm of LDA consists of the following steps:

- For each class $i = 1 \ldots C$, calculate mean μ_i, covariance matrix Σ_i and *a priori* class probability $P_i = \frac{N_i}{N}$, where N_i is the number of vectors in class i.
- Calculate the mean of the entire data set (includes all classes):
  - $\mu = \sum_{i=1}^{C} P_i \mu_i$
- Calculate scatter matrices:
  - $S_W = \sum_{i=1}^{C} P_i \Sigma_i$
  - $S_B = \sum_{i=1}^{C} P_i (\mu_i - \mu)(\mu_i - \mu)^T$
- Do eigen-analysis of $S_W = VDV^T$
- Discard any zero eigen-values of Sw with their eigen-vectors
- Form whitening transform $B = VD^{-1/2}$
- Obtain new decorrelated $S'_B = B^T S_B B$
- Do eigen-analysis of $S'_B = U\Lambda U^T$
- Select M biggest eigen values $\lambda_1 \ldots \lambda_M$
- Pack associated eigen vectors into transformation matrix $W = B \times [u_1 \ldots u_M]$
  - W rotates and scales the data
- Transform data via $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$, where **X** is the original data matrix and **Y** is the transformed data matrix
  - Maximum number of dimensions after LDA is C - 1

From the transformed data, classification is typically performed using the Euclidean distance or RMS distance of the unknown case to the centroids of the class sets.

The usefulness of LDA are:

- Between-class separatebility is maximized
- Within-class differences are reduced
- Variables of the data set are decorrelated

The drawbacks and limitations of LDA are:

- Dependent only on means and variance
    - Favors uni-modal data distributions
- Outliers may distort results
- May miss nonlinear structures

Appendix H: Data for Misclassified /d/ and /ð/ Cases

	Fhi	**AhiA2**	**AhiVhi**	**F2**	**Cog**	**Stdv**	**Skew**	**Kurtosis**
μ	3730Hz	2.5dB	3.9dB	1800Hz	4200Hz	1370Hz	0.79	0.61
σ	1060Hz	3.8dB	4.3dB	300Hz	630Hz	200Hz	0.67	1.6
1	4639	-0.87687	7.078888	***1276.768***	***4867.576***	1468.761	0.13709	-0.29207
2	***2004***	0.001787	6.269794	1853.071	4702.714	***1646.042***	***-0.12121***	***-1.11308***
3	***6296***	***-2.92894***	***10.65507***	2057.863	***5090.662***	***1854.607***	***-0.15645***	***-1.47901***
4	3628	-0.86539	-0.05695	1971.604	4026.883	***1781.285***	1.1269	-0.05179
5	***2432***	***-1.30672***	0.09578	1553.21	***5022.563***	***1785.547***	***-0.00028***	***-1.28729***
6	***7678***	***13.31033***	***11.32932***	2006.141	***6402.248***	1504.762	***-1.2201***	0.477896
7	***2157***	2.395291	0.073018	***2184.861***	4696.197	***1700.47***	***-0.04555***	***-1.26559***
8	4375	0.49637	3.525665	1784.684	4330.262	***1614.343***	0.628883	-0.55771
9	***5626***	-0.07264	5.253238	2008.077	4749.41	***1793.662***	0.053887	***-1.27381***
10	3856	-0.25491	2.359716	1717.881	4580.273	***1665.887***	0.521365	-0.92092
11	***5218***	1.732003	3.904976	1610.2	4519.697	1455.616	***0.057074***	-0.57443
12	3774	0.058668	4.849631	1598.1	4518.87	***1571.056***	0.603251	-0.6002
13	***2418***	0.26084	4.089787	***1462.519***	4036.378	***1661.343***	0.66587	-0.78479
14	***4877***	***-1.82287***	5.803072	1828.913	4829.8	***1682.314***	0.24377	-0.96562
15	***2698***	-0.66278	***-2.86725***	1880.707	4390.803	***1718.171***	0.540757	-0.75039
16	4589	***-1.54176***	1.829729	***2171.587***	***5013.171***	***1707.194***	***0.038297***	***-1.20827***
17	***6254***	2.285235	7.932532	1946.066	4755.113	1520.974	***0.063795***	-0.93939
18	4543	-0.31106	***-1.05774***	1875.391	4475.906	1535.416	0.131222	-0.96583
19	***2509***	***-2.95016***	***9.918222***	***1492.395***	4318.94	1502.411	0.363398	-0.80034
20	***2411***	-0.7763	6.040689	***1095.124***	4242.054	***1696.889***	0.647038	-0.62252
21	***6158***	***10.28546***	4.798669	1947.223	***5730.156***	1282.828	***-1.12458***	0.73403

Table H1: Data for misclassified /d/ tokens. The table contains the mean (μ) and standard deviation (σ) for each variable across all 102 /d/ tokens, as well as the data for the 21 misclassified /d/ tokens.

	Fhi	AhiA2	AhiVhi	F2	Cog	Stdv	Skew	Kurtosis
μ	4480Hz	0.22dB	2.5dB	1670Hz	4860Hz	1640Hz	0.08	-0.91
σ	1830Hz	3.3B	6.2dB	330Hz	480Hz	170Hz	0.38	0.49
1	***2233***	-0.97589	8.491591	***2349.93***	4474.046	1634.181	0.361735	-0.90397
2	***7304***	2.973191	***10.7418***	1702.221	***6069.171***	1533.202	***-1.13438***	***0.217052***
3	5866	***5.192809***	0.172997	1771.412	***5543.887***	***1376.201***	-0.2378	-0.65212
4	***2151***	2.235841	-2.09052	***1292.243***	***4199.016***	***1831.163***	0.231311	***-1.4086***
5	3789	0.405662	6.500049	***2263.261***	***4198.027***	1331.015	***0.605169***	-0.08078
6	2891	0.695276	3.661339	***1212.143***	***4153.46***	***1413.546***	***0.72118***	-0.04408
7	3989	-2.10085	6.38799	1451.304	***3915.366***	***1423.934***	***0.970626***	***0.269675***
8	4088	2.035517	***9.897998***	1416.296	4514.227	1521.681	***0.531238***	-0.73095
9	2912	***4.068034***	3.924127	1588.311	4757.373	1663.204	0.228984	-1.26969
10	2772	***4.158777***	6.881437	1833.898	4546.663	1507.023	0.317054	-0.8862
11	***2380***	-2.49753	***-3.79417***	1484.231	***3425.257***	***1433.91***	***1.244146***	***0.626045***
12	3683	3.756052	-0.50417	***2202.4***	4412.123	1596.407	0.412026	-0.80518
13	5310	***7.732915***	8.272745	***2259.019***	4630.618	1490.818	0.178358	-0.75915
14	3474	***5.562867***	***8.728799***	***2282.764***	***4093.117***	***1202.305***	***1.009072***	***0.863856***
15	4356	-0.82706	2.901399	***2050.9***	4672.083	1474.557	0.301135	-0.57984
16	5389	***12.87104***	***9.444866***	***2186.389***	***5792.063***	***1143.735***	***-0.39348***	***0.874027***
17	5253	1.37564	***9.126374***	1400.872	5259.705	***1251.497***	***-0.33089***	***0.183622***
18	3900	2.263331	7.089904	1578.123	***4268.05***	1526.37	***0.579078***	-0.58733

Table H2: Data for misclassified stop-like /ð/ tokens. The table contains the mean (μ) and standard deviation (σ) for each variable across all 100 /ð/ tokens, as well as the data for the 18 mis-classified /ð/ tokens.

Appendix I: Matlab Source Code for Automatic Classification

```
function AutoClass8

[d_Fhi, dh_Fhi]=groupAB('train5_d_F_hi.txt', 'train5_dh_F_hi.txt');
[d_Ahi2Alo, dh_Ahi2Alo]=groupAB('train5_d_ahi_to_a2.txt', 'train5_dh_ahi_to_a2.txt');
[d_F2, dh_F2]=groupAB('train5_d_Formant2.txt', 'train5_dh_Formant2.txt');
[d_AhiVhi, dh_AhiVhi]=groupAB('train5_d_ahi_to_vhi.txt', 'train5_dh_ahi_to_vhi.txt');
[d_Cog, dh_Cog]=groupAB('train5_d_cog.txt', 'train5_dh_cog.txt');
[d_sdv, dh_sdv]=groupAB('train5_d_sdvf.txt', 'train5_dh_sdvf.txt');
[d_skew, dh_skew]=groupAB('train5_d_skewness.txt', 'train5_dh_skewness.txt');
[d_kurt, dh_kurt]=groupAB('train5_d_kurt.txt', 'train5_dh_kurt.txt');

d_len = length(d_Fhi);
dh_len = length(dh_Fhi);

training_Fhi=[d_Fhi' dh_Fhi']';
training_Ahi2Alo=[d_Ahi2Alo' dh_Ahi2Alo']';
training_F2=[d_F2' dh_F2']';
training_AhiVhi=[d_AhiVhi' dh_AhiVhi']';
training_Cog=[d_Cog' dh_Cog']';
training_Sdv=[d_sdv' dh_sdv']';
training_Skew=[d_skew' dh_skew']';
training_Kurt=[d_kurt' dh_kurt']';


combined_train=[training_Fhi'; training_F2'; training_Ahi2Alo'; training_AhiVhi'; training_Cog';
training_Sdv'; training_Skew'; training_Kurt']';

groups(1:d_len)=1;
groups(d_len+1:d_len+dh_len)=2;

[coeff score] = princomp(zscore(combined_train));
coeff,

[coeff lat explain] = pcacov(cov(zscore(combined_train))),


%*************** LEAVE-ONE-OUT/CROSS-VALIDATE CLASSIFY ****************%

group_sz = length(groups);
class_mahal=zeros(group_sz, 1);
class_lda=zeros(group_sz, 1);

for (count=1:group_sz)
   test_data = score(count, :);
   if(count == 1)
      training_data = score(count+1:group_sz, :);
      groupsN = groups(count+1:group_sz);
   elseif(count == group_sz)
      training_data = score(1:group_sz-1, :);
      groupsN = groups(1:group_sz-1);
   elseif(count == 2)
      training_data = [score(1,:); score(count+1:group_sz, :)];
```

```
      groupsN = [groups(1) groups(count+1:group_sz)];
   elseif(count == group_sz-1)
      training_data = [score(1:count-1,:); score(count+1, :)];
      groupsN = [groups(1:count-1) groups(count+1)];
   else
      training_data = [score(1:count-1,:); score(count+1:group_sz, :)];
      groupsN = [groups(1:count-1) groups(count+1:group_sz)];
   end
   % since the PCs are independent and have a diagonal covariance matrix, the mahalanobis
    % distance is the same as normalized Euclidean distance (normalized by variance of PCs)
     class_mahal(count)=classify(test_data, training_data, groupsN, 'mahal');
     %class_lda(count)=classify(test_data, training_data, groupsN, 'linear');
end

class_mahal,

for (count=1:group_sz)
   test_data = combined_train (count, :);
   if(count == 1)
      training_data = combined_train(count+1:group_sz, :);
      groupsN = groups(count+1:group_sz);
   elseif(count == group_sz)
      training_data = combined_train(1:group_sz-1, :);
      groupsN = groups(1:group_sz-1);
   elseif(count == 2)
      training_data = [combined_train(1,:); combined_train(count+1:group_sz, :)];
      groupsN = [groups(1) groups(count+1:group_sz)];
   elseif(count == group_sz-1)
      training_data = [combined_train(1:count-1,:); combined_train(count+1, :)];
      groupsN = [groups(1:count-1) groups(count+1)];
   else
      training_data = [combined_train(1:count-1,:); combined_train(count+1:group_sz, :)];
      groupsN = [groups(1:count-1) groups(count+1:group_sz)];
   end
   class_lda(count)=classify(test_data, training_data, groupsN, 'linear');
end

class_lda,

. . . . . .

function [sampleA, sampleB]=groupAB(fnameA, fnameB)

fidA=fopen(fnameA, 'r');
fidB=fopen(fnameB, 'r');

sampleA=fread(fidA, 'double=>double');
sampleB=fread(fidB, 'double=>double');

%sampleA=listA;
%sampleB=listB;

fclose(fidA);
fclose(fidB);
```