Multimodal Speech Recognition with Ultrasonic Sensors

by

Bo Zhu

S.B., Massachusetts Institute of Technology (2007)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2008

Author . . . .
Department of Electrical Engineering and Computer Science
July 23, 2008

Certified by . . . .
James R. Glass
Principal Research Scientist
Thesis Supervisor

Certified by . . . .
Karen Livescu
Research Assistant Professor
Thesis Co-Supervisor

Accepted by . . . .
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

Multimodal Speech Recognition with Ultrasonic Sensors

by

Bo Zhu

Submitted to the Department of Electrical Engineering and Computer Science
on July 23, 2008, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Ultrasonic sensing of articulator movement is an area of multimodal speech recognition that has not been researched extensively. The widely-researched audio-visual speech recognition (AVSR), which relies upon video data, is awkwardly high-maintenance in its setup and data collection process, as well as computationally expensive because of image processing. In this thesis we explore the effectiveness of ultrasound as a more lightweight secondary source of information in speech recognition.

We first describe our hardware systems that made simultaneous audio and ultrasound capture possible. We then discuss the new types of features that needed to be extracted; traditional Mel-Frequency Cepstral Coefficients (MFCCs) were not effective in this narrowband domain. Spectral analysis pointed to frequency-band energy averages, energy-band frequency midpoints, and spectrogram peak location vs. acoustic event timing as convenient features.

Next, we devised ultrasonic-only phonetic classification experiments to investigate the ultrasound's abilities and weaknesses in classifying phones. We found that several acoustically-similar phone pairs were distinguishable through ultrasonic classification. Additionally, several same-place consonants were also distinguishable. We also compared classification metrics across phonetic contexts and speakers.

Finally, we performed multimodal continuous digit recognition in the presence of acoustic noise. We found that the addition of ultrasonic information reduced word error rates by 24-29% over a wide range of acoustic signal-to-noise ratio (SNR) (clean to 0dB). This research indicates that ultrasound has the potential to be a financially and computationally cheap noise-robust modality for speech recognition systems.

Thesis Supervisor: James R. Glass
Title: Principal Research Scientist

Thesis Co-Supervisor: Karen Livescu
Title: Research Assistant Professor

Acknowledgments

I would first like to thank my advisor, James Glass, who has provided me with immense support and patience over the past few years. His encouragement during tough times and advice on a variety of issues were invaluable. I would like to thank my co-advisor, Karen Livescu, for her continual guidance and wisdom. Working alongside her during experiments was a joy and a privilege.

I also thank T.J. Hazen, who helped me tremendously with the digit recognizer, and I would like to thank Lee Hetherington for all his assistance with Sapphire. Much thanks to Carrick Detweiler and Iuliu Vasilescu for their work on the next generation hardware capture device and helping me set up for data collection. I also would like to thank Hung-An Chang for his help on the classifiers, as well as Ken Schutte, Paul Hsu, and Ghinwa Choueiter for their constant willingness to help with any question I have.

My deep gratitude goes out to my friends from MIT and high school, who have encouraged and supported me along the way. Of course, I would not be here without my parents, to whom I cannot thank more for their love and sacrifices. Finally, special thanks to my brothers, Ted and Tim, whom I become more proud of every day.

This research was supported by Quanta Computer.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Conventional automatic speech recognition (ASR) systems use only audio information. When the speech audio becomes corrupted by the presence of external noise, recognition performance suffers.

There are three main ways to deal with channel noise. One is to do audio preprocessing on the noisy signal in order to recover as much meaningful data as possible. This might involve methods such as adaptive filtering and spectral subtraction. The second is to use model-based techniques to model the speech+noise signal (e.g. [5]). The third is to simultaneously use another sensor that will capture the same linguistic information but in another domain, often called the multi-modal approach. When the noise is non-stationary, which includes babble speech noise, the first method usually performs poorly [24]. In this research, we explore the use of ultrasonic sensors aimed at the user's mouth. These sensors obtain information corresponding to movements around the lower facial region while operating at frequencies beyond the audible range. Thus environmental noise will not affect these sensors, and this clean source of secondary information will add noise robustness to the ASR system.

1.1 Motivation

Humans regularly perform multimodal speech recognition. Watching someone speak allows one to gather information about place of articulation and audio source localization [20]. Therefore, vision as a secondary information source is useful for speech recognition in a loud environment. Researchers have translated this to automatic speech recognition by using video cameras to capture facial features during speech (audio-visual speech recognition, or AVSR), with significant improvements in recognition performance [19]. However, high-resolution video cameras can be quite expensive, and the image processing and high dimensionality of data used in classification can also be computationally expensive. The physical limitations (described in Related Work in Chapter 2) also make current AVSR setups impractical for the average consumer.

In searching for cheaper (both materially and computationally) alternatives, researchers have tested a multitude of sensors ranging from "tethered" skin-conducting microphones [24, 7, 13] to "untethered" [18, 10, 11] sensors operating from a distance. These multimodal systems will be described in further detail in the Related Work chapter.

In 2006, Kalgaonkar and Raj showed that by using ultrasonic sensors, effective multimodal voice activity detection could be done [11]. More importantly, they established that certain aspects of lip movement could be quantified by the Doppler effect and measured by frequency changes in the ultrasonic signal. Inspired by this experiment, we set out to extend the use of these sensors to perform speech recognition tasks.

1.2 Ultrasonic sensors background

Ultrasonic transducers are constructed from piezoelectric materials (usually ceramic) that bend at a set resonant frequency above 20kHz. There are two types of ultrasonic transducers: transmitters and receivers. Transmitters radiate inaudible sound waves given an input voltage, while receivers output voltage given the received sound waves. A transmitter/receiver pair is shown in Figure 1-1.

Ultrasonic sensors are used for a variety of applications, such as rangefinding [9] and medical imaging [15], and have different transmitter/receiver setups. Addition-

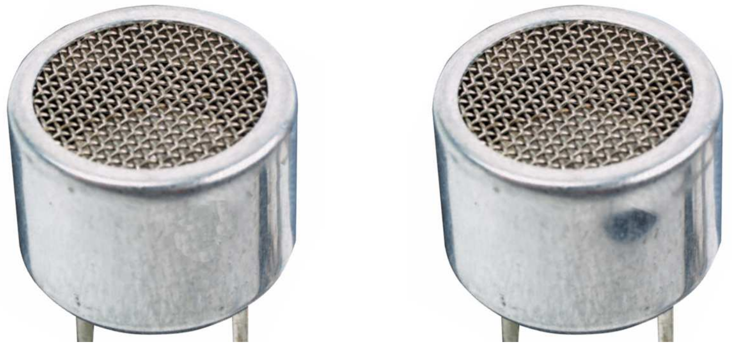


Figure 1-1: Ultrasonic transmitter and receiver

ally, they are driven in different ways. Ultrasonic rangefinders usually send a periodically pulsed signal to a single transmitter. The pulse reflects off an object, and the time it takes to get back to the receiver can thus be measured. For our setup, we transmitted a continuous sinusoidal signal because we were interested in the continuous frequency response governed by the Doppler effect.

The Doppler effect states that the frequency observed by the receiver sensor depends upon the velocity of the sound source and the frequency emitted by that source. In equation form:

$$f = f_0(1 + \frac{v}{c}) \tag{1.1}$$

where f is the observed frequency, f_0 is the frequency at the sound source, v is the velocity of the sound source in the direction of the receiver, and c is the constant speed of sound. Therefore, a constant frequency emitted by an object moving toward the receiver will result in an observed increase in frequency, while the same object moving away from the receiver will result in a decreased frequency (e.g. police sirens increasing or decreasing in pitch as the car moves towards or away from the observer, respectively).

This phenomenon can be shown with a simple experiment. We direct an ultrasonic transmitter (driven at 40 kHz) at a sheet of cardboard. We move the cardboard toward and away from the transmitter/receiver pair with its face parallel to the transducers. In the reference frame of the receiver, the cardboard is the sound source, not the transmitter. Therefore, movement of the cardboard will change v in

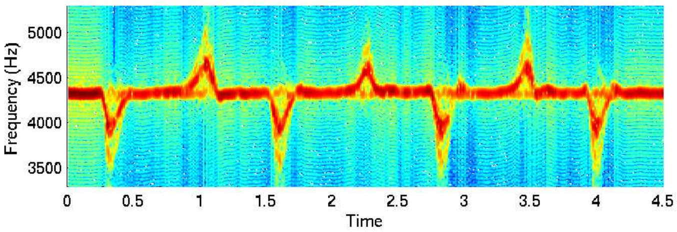


Figure 1-2: Example of Doppler Effect. Shown above is a spectrogram of the ultrasonic signal of cardboard being pushed and pulled away from the transmitter/receiver pair. Cardboard movement causes frequency shifting in accordance to the Doppler effect.

Equation 1.1. Because the transmitted frequency is constant, the observed frequency f will be directly correlated to the cardboard velocity v. The spectrogram of the received ultrasonic signal is shown in Figure 1-2. The frequency changes are a result of cardboard movement. Note that the signal was downsampled during postprocessing; the frequency scale shown in this figure is much lower than the actual 40 kHz operating range of the experiment.

To apply the Doppler effect to speech recognition, we direct a transducer/receiver pair toward a user's mouth. The lower facial region can be modeled as a mesh of infinitely small reflecting surfaces. Each surface reflects a 40 kHz signal toward the receiver, and moves independently of the other surfaces. The result at the receiver is a superposition of sinusoids at different frequencies. This can be seen in Figure 1-3 below, which are ultrasonic and audio spectrograms of a user speaking the utterance "ma na". From the ultrasonic spectrogram, we can see that there are indeed many frequencies represented in each frame.

Additionally, from this figure we can begin to see the value of an ultrasonic sensor towards the application of speech recognition. The nasals "m" and "n" are traditionally difficult to distinguish in standard ASR systems because of their concentration of energy in the low frequency range. One can see the overwhelming similarities in the audio spectrogram: almost all the energy is packed into the lowest frequencies. However, the articulatory motions for these sounds "ma" and "na" are very different, and this is evidenced by the clear differences in the ultrasonic spectrogram.

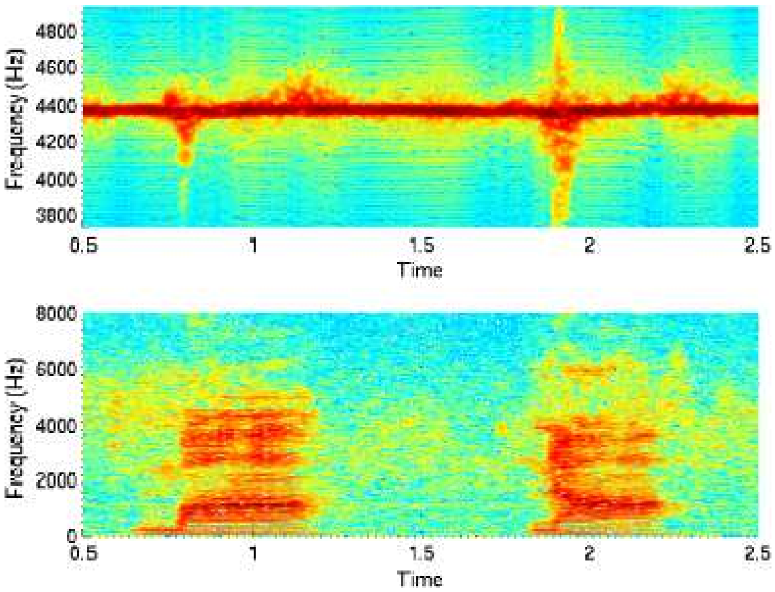


Figure 1-3: Ultrasonic (top) and audio (bottom) spectrograms of a user speaking "ma na". Nasals that are acoustically difficult to distinguish are easily differentiable in the ultrasonic spectrogram.

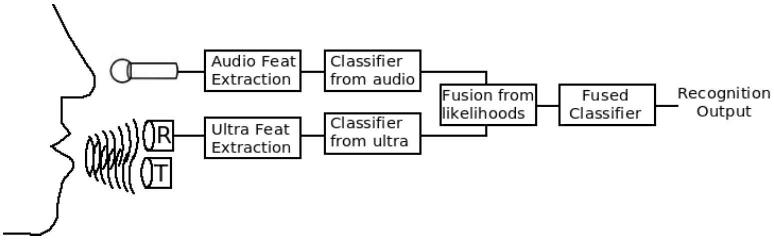


Figure 1-4: Block diagram of multimodal ASR system

1.3 Proposed approach

The purpose of this project is to study the effectiveness of ultrasonic signals as a secondary mode of information in the context of automatic speech recognition. The proposed system is shown in Figure 1-4.

Physically, hardware must be created to transmit and receive ultrasonic signals. Detailed descriptions of our hardware will follow in the next section. The hardware outputs the ultrasonic (as well as microphone audio) signal to a stereo minijack cable.

The cable is connected to a computer's sound card for input. The signals are processed in software; most notably, feature extraction is performed on the raw audio and ultrasonic signals to select only the most relevant aspects of the data for classification.

The ultrasonic signals are fundamentally different than the audio signals; they are narrowband and frequency-modulated around a carrier, so the traditional wideband Mel-Frequency Cepstral Coefficients (MFCCs) will not be appropriate as ultrasonic features. New methods of feature extraction must be created for this ultrasonic data. The specific methods of feature extraction that we developed are detailed below in Chapter 4.

Using these feature extraction methods, models are learned from newly collected data for recognition of separate test data. We perform noisy digit recognition experiments, varying both the acoustic noise level and the weights given to the audio and ultrasonic models in the scoring process. Results from these experiments provide insight into the effectiveness of the ultrasonic information in improving noisy speech recognition.

Additionally, we would like to investigate the ultrasound's ability to classify between specific phonemes, phoneme groups, and articulatory motions. We therefore experiment with a separate set of data consisting of Consonant-Vowel-Consonant (CVC) and Vowel-Consonant-Vowel (VCV) utterances. We derive models from the ultrasonic features and perform classification experiments.

1.4 Overview

The remainder of this thesis is organized in the following manner. Chapter 2 will discuss previous related work in the areas of multimodal speech recognition, multimodal fusion techniques, and ultrasonic speech processing and recognition. Chapter 3 will describe our hardware setup used to capture the acoustic and ultrasonic signals for data collection. We will present a prototype system as well as a newer hardware setup. Chapter 4 will detail the feature extraction methods used to transform the acoustic and ultrasonic data into feature vectors for classification and recognition. We begin testing the effectiveness of the ultrasound with Chapter 5, which will describe the ultrasonic phonetic classifier along with results. Chapter 6 will describe our digit recognizer and the implications of our findings. We will conclude with Chapter 7.

Chapter 2

Related Work

2.1 Multimodal speech recognition

It has been known for a while now that humans integrate multiple sources of information to recognize speech [20, 14]. The most popular secondary source of information is visual. In a noisy environment, humans use lip reading as well as facial expressions and gestures. In fact, the interesting phenomenon of superadditivity often occurs: The accuracy of speech perception with two information sources is often greater than the sum of the accuracy measures of the individual sources [3, 2].

Speech recognition scientists have been taking advantage of multimodal perception, using it for noise-robust machine speech recognition. Zhang et al. [24] at Microsoft Research have produced prototypes that rely upon secondary information from bone-conducting sensors. The bone-conductive microphone captures speech in the $<$ 3kHz frequency range. The data from this sensor is used for voice activity detection as well as estimating a reconstruction of the original clean waveform. Voice activity detection (VAD) simply outputs whether or not the user is speaking. This is useful in reducing recognition of noise as nonsense speech. Additionally, the estimated reconstruction of the clean speech signal is fed into a speech recognizer. The published speech recognition experiment was speaker dependent and performed on a 42-sentence corpus, and they showed that their WER was reduced from 64% to 30% when using the additional bone-conducting signal.

Graciarena et al. [7] have used a glottal electromagnetic micropower sensor (GEMS), which is attached to the skin near the user's throat. The sensor is an extremely sensitive phase-modulated quadrature motion detector. The probabilistic optimum filter (POF) method [17] is used to map the noisy microphone + throat microphone Mel-Frequency cepstral coefficient (MFCC) features to clean MFCC features. POF is an implementation of feature concatenation fusion, which will be discussed below in Section 2.2. This multimodal system yielded a WER reduction from 95.6% to 52.6% in 0dB SNR. Kwan et al. have also made a GEMS multimodal speech recognition system. They used Gaussian mixture models instead of POF to reconstruct the clean MFCC features. This yielded a WER reduction from 60% to 40% at 5dB SNR [13].

The most widely researched modality of multimodal speech recognition is Audio-Visual Speech Recognition (AVSR). For humans, information about place of articulation is obtained when looking at the speaker's mouth, which increases human speech recognition performance [19]. Applied to machines, AVSR uses a video camera to capture visual information about the user's face and an acoustic microphone to capture simultaneous audio data from the speech. Processing of the visual information results in visual features that are ultimately fused with acoustic features and fed into a recognizer that takes into account both types of information. State-of-the-art AVSR systems are able to reduce WER from 78% to 38% in 0dB SNR [19]. However, AVSR requires computationally expensive preprocessing just to prepare the data: face and facial part recognition, region of interest (ROI) extraction, (optionally) lip/face shape recognition, lighting normalization. There are also stringent physical limitations, such as no side-to-side rotation of the head, and stationary, no-shadow light source location. The two-dimensional nature of the video images also result in computationally expensive processing during the actual feature extraction and later stages.

2.2 Fusion of multimodal sources

For two information sources to be considered in the recognition of speech, there must be a way to fuse these sources to obtain one recognition result in the end.

Fusion from features, also known as Direct Identification (DI) fusion [19], usually require either feature concatenation of two sources or feature weighting. In feature concatenation, each feature source is treated as equally important, and the two are simply combined in one long feature vector. Usually dimensionality reduction is necessary afterwards [16]. This single vector is then used in the classification stages. In feature weighting, the audio or secondary-sensor Gaussian distance-to-mean of each model is multiplied by a certain weight. This allows flexibility in choosing the contribution level of each modality. Fusion from classifier outputs is known as Separate Identification (SI). SI is very good at taking advantage of the reliability of each modality [19]. SI fusion is usually done by a linear combination of log-likelihoods of each single-modality score. The linear combination uses defined weights for each modality [22]. These fused log-likelihoods are then used to determine the output. There is also active research in automatically finding the optimum weight for each modality [21]. There exist other more complex fusion methods, particularly those which use both DI and SI. These are known as Hybrid fusion techniques, which can generally perform better than DI- or SI-only methods [19, 4, 8].

2.3 Ultrasonic speech recognition modality

There has not been much research on using ultrasonic sensors as a second modality in speech recognition. Jennings and Ruck [10] created a multimodal ultrasonic speech recognition system based on dynamic time warping (DTW), with experiments on speaker-dependent, isolated digit recognition.

In their setup, they used a 40 kHz oscillator to drive an ultrasonic transducer aimed at the user's mouth. The signal reflects back, and a standing wave manifests between the transmitter/receiver pair and the user's mouth. Mouth movements change the amplitude and slightly shift the frequency of the standing wave. This ultrasonic signal is captured by the ultrasonic receiver, and is fed through an envelope detector and AC coupling. This low-frequency signal is downsampled and used as features in the ultrasonic classifier. The acoustic features are 10 LPC coefficients per frame.

For each class, a template is defined for each modality, from which DTW distances are derived. The probability of a certain class (for each modality) is inversely proportional to the DTW distance and is normalized over the distances for each class. These “pseudo probability mass functions” for each modality are fused pairwise by a simple linear combination, resulting in one output probability for each class. Jennings and Ruck performed speaker-independent experiments of isolated digit recognition, adding various levels of white noise to the acoustic channel. At 0dB, the system was able to reduce WER from 22% to 7%.

More recently, Kalgaonkar and Raj [11] used a similar hardware setup to perform voice activity detection using a multimodal ultrasonic system. Changes in mouth movement are characterized by Doppler frequency shifts. The detection algorithm frequency-demodulates the received signal, and the energy of the resultant signal represents the total velocity of the articulators. This energy is compared to an adaptive threshold, and the output is a binary “speech” or “no speech” decision. With 0dB babble noise, VAD detection rate increased from 52.5% audio-only to 96.05% using both audio and ultrasonic information.

Chapter 3

Hardware Setup

3.1 Prototype hardware

The hardware is the first part of the multimodal system, capturing the ultrasonic and acoustic information simultaneously. The ultrasonic part needs to generate and receive an ultrasonic signal. The acoustic signal is captured by a simple electret microphone. This section describes the first hardware capture system we built.

Each of the received ultrasonic and acoustic signals are single-channel, so we can output them as a stereo signal, using a stereo minijack cable. This cable is input to a conventional computer sound card, which performs A/D conversion. Since a 40 kHz carrier tone is higher than the largest sampling rate of most sound cards, we decided to modulate the ultrasonic signal so that the stereo signals could be sampled without aliasing at a sampling frequency of 16 kHz/s. The hardware setup is shown in Figure 3-1.

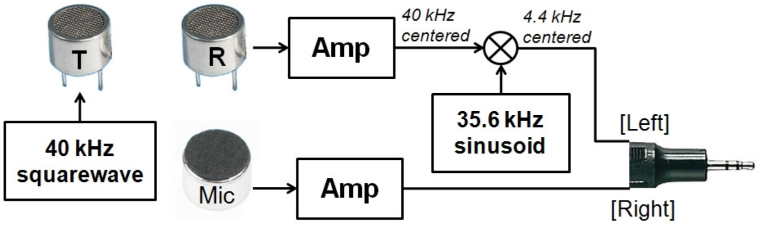


Figure 3-1: Hardware configuration

The transducers it contains are an ultrasonic emitter and receiver, and an electret microphone for the regular audio signal. The ultrasonic transmitter is a Kobitone 400ST160 tuned to a resonant frequency of 40 kHz. The transmitter is driven by a 40 kHz squarewave generator, which is implemented by a PIC10F206 microcontroller. The output of the transmitter is a pure sinusoid even though it is driven by a squarewave, because the transmitter is inherently a narrowband device that will bandpass filter the other harmonics, leaving the first 40 kHz sinusoidal harmonic.

The ultrasonic receiver is a Kobitone 400SR160 also centered around 40 kHz, with a -6dB bandwidth of 2.5 kHz. This receiver is extremely sensitive within this bandwidth, which allows minor frequency shifts to be detected; these frequency shifts are the basis of our subsequent analysis. In order to shift the ultrasonic spectrum down to a lower frequency range, the received signal is modulated with a 35.6 kHz sinusoid to downshift it to be centered at 4.4 kHz, well within the capture bandwidth of standard sound cards. The modulation process is implemented by a 35.6 kHz squarewave generator (also a PIC10F206) and a fourth-order Butterworth lowpass filter with a cutoff frequency at 48 kHz. This cutoff frequency eliminates the odd harmonics above the first, resulting in a 35.6 kHz sinusoid. An Analog Devices MLT04 analog multiplier is then used to multiply the received signal and the sinusoid to perform the modulation.

The schematics and Printed Circuit Board (PCB) layout are shown in Figures A-1 and A-2 in Appendix A. The PCB was printed offsite by a PCB manufacturing company and sent back to us. We hand-soldered the device with the necessary components onto the PCB and tested it to ensure that the hardware was working properly.

3.2 Next-generation hardware

In order to reduce noise in the capture process, as well as to miniaturize our current setup, we collaborated with researchers Carrick Detweiler and Iuliu Vasilescu from the MIT CSAIL Distributed Robotics Laboratory to build a small, digital-output version of the ultrasonic+audio capture hardware.

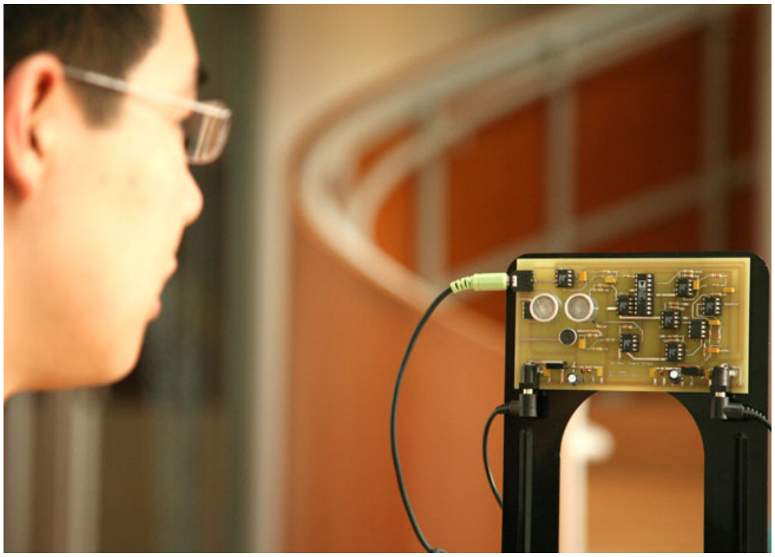


Figure 3-2: User speaking into prototype hardware

Figure 3-4 shows an image of the next-generation hardware and Figure 3-3 shows its block diagram. At the heart of the device is a Xilinx XC2C256 CoolRunner II CPLD (complex programmable logic device). This generates a 40 kHz square wave with variable duty cycle which is input into the ultrasonic emitter.

The reflected signal is captured by the ultrasonic receiver. This signal passes through a low noise amplifier (LNA) followed by a variable gain amplifier (VGA), which allows control over the sensitivity of the receiver (36dB range). The signal then passes through a 40 kHz bandpass filter. Finally, the filtered signal goes into a 16-bit analog to digital converter (Analog Devices AD7680). The CPLD reads the ADC at 24 kHz causing the 40kHz signal from the ultrasonic receiver to be aliased down to 8kHz.

The audio is captured from an internal or external microphone and is processed similarly to the ultrasound channel. The main difference is that a lowpass filter with a cutoff of around 8 kHz is used. The digital representation of both the ultrasound and audio channels on the CPLD is then formatted for transfer over USB using an FTDI FT245 USB chip. The result is pure digital streams of both channels to the host computer.

The fabrication process was handled by Carrick and Iuliu, and reportedly was

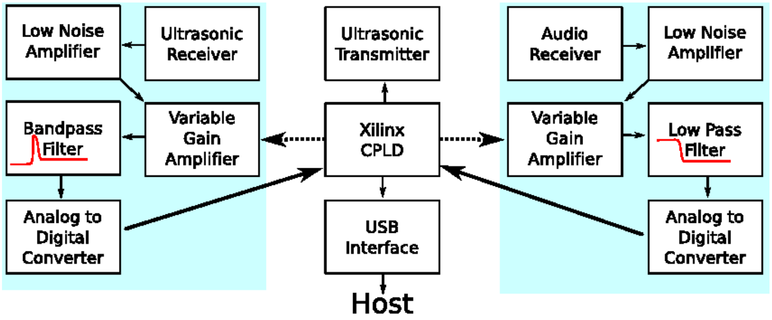


Figure 3-3: Block diagram of next-generation hardware system

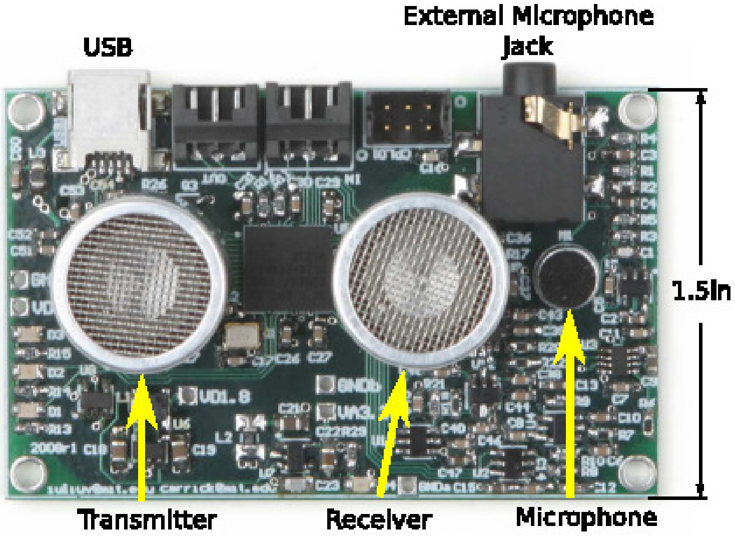


Figure 3-4: Next-generation hardware

similar to the process for the prototype board.

Chapter 4

Feature Extraction

The audio and ultrasonic channels will each go through independent feature extraction stages, whose outputs will be used in separate classifiers. The audio channel will be processed by standard Mel-Frequency Cepstral Coefficient (MFCC) feature extraction, while new techniques must be developed to extract features from the fundamentally different type of data in the ultrasonic channel.

4.1 Audio Feature Extraction

We use standard MFCCs as features from the audio signal. MFCCs represent the spectral content of a signal on a logarithmic frequency scale.

The audio signal is split into 5 ms frames, and the spectrum is calculated by Fast Fourier Transform (FFT) analysis. The FFT coefficients are then mapped to a logarithmic "Mel-scale" using triangular windows, shown in Figure 4-1[1]. The log-power at each of these mel frequencies is calculated, and then the Discrete Cosine Transform (DCT) of this log-power spectrum is computed, resulting in the MFCCs [1]. We extract 14 total MFCC features from each frame.

[1] http://labrosa.ee.columbia.edu/doc/HTKBook21/img165.gif

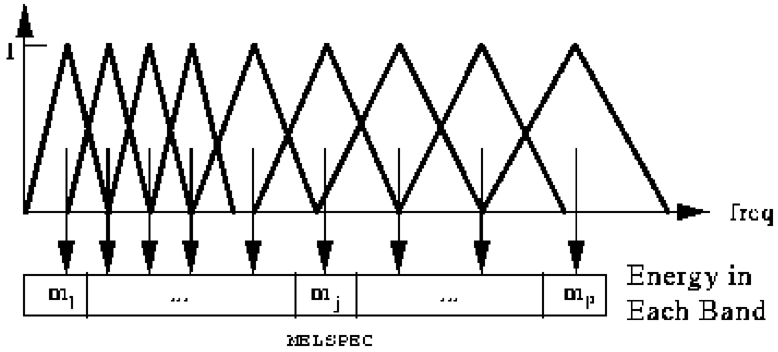


Figure 4-1: Illustration of the Mel-scale and triangular averaging windows

4.2 Ultrasonic Feature Extraction

The ultrasonic signal differs from the acoustic signal in that it is narrowband and very sensitive to minute frequency shifts. Standard MFCC features are not sufficient for classification. We must analyze this ultrasonic signal and develop a new method of feature extraction.

4.2.1 Time domain analysis

Figure 4-2 shows audio spectrogram and ultrasonic time-domain plots of four utterances: two of the digit “seven” and two of the digit “five”. The utterances were taken from continuous speech sequences in which the instances of “seven” and “five” occurred.

Across instances of “seven” and “five”, we observe that the ultrasonic plots show almost no correlation whatsoever between digits across instances. Subsequent analysis of other digits across more instances reveals a similar trend. There does not seem to be a robust, quantifiable measurement that will allow reasonable classification to be done. Thus we must look elsewhere to gain insight for ultrasonic feature extraction.

4.2.2 Frequency domain analysis

As described earlier, the recorded ultrasonic signal will consist of a number of different frequency components, with each component corresponding to a reflection from a moving (articulator) surface. The amount of energy at a particular frequency can be

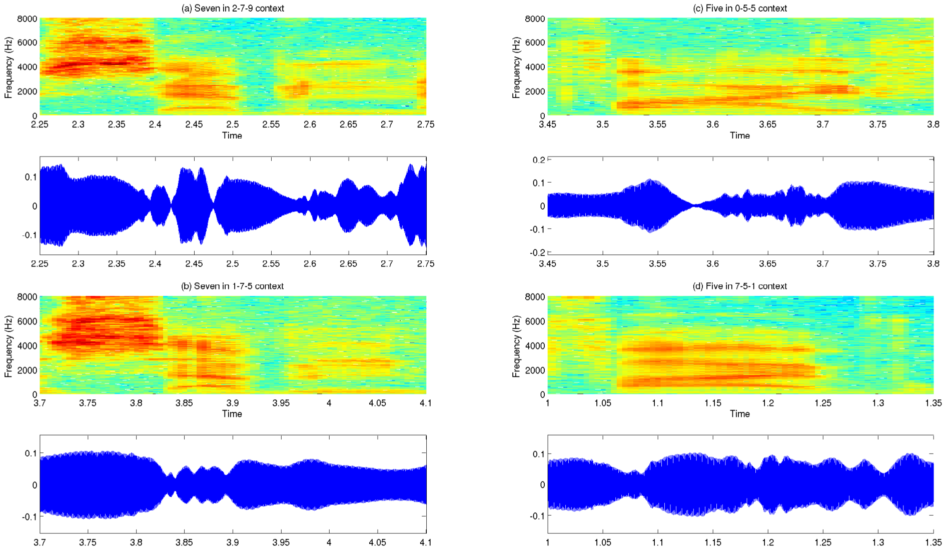


Figure 4-2: Time-domain ultrasonic analyses of two instances each of “seven” (a,b) and “five” (c,d). There is seemingly very little ultrasonic correlation between (a) and (b), as well as (c) and (d).

associated with articulator(s) moving with a certain velocity at a particular time. We can thus expect the spectrograms of identical utterances to appear similar.

This is confirmed in Figure 4-3, which shows the same utterances as those in Figure 4-2, but substitutes ultrasonic time-series plots with spectrograms. We can observe much greater similarities between instances of the same digit. We would like to extract features from the spectra of the ultrasonic signals.

4.2.3 Carrier cancellation

In addition to the ultrasonic reflections from the user, the receiver also picks up coupling directly from the transmitter. We can see from the middle graph in Figure 4-4 that the carrier signal is very strong, and in fact it overwhelms the magnitudes of the ultrasonic signal near the carrier. We would like to remove this coupled signal by spectral subtraction.

We characterize the carrier spectrum by taking the FFT of the first 6 ms frame of each utterance, when there is no movement or talking. Figure 4-5 shows a typical car-

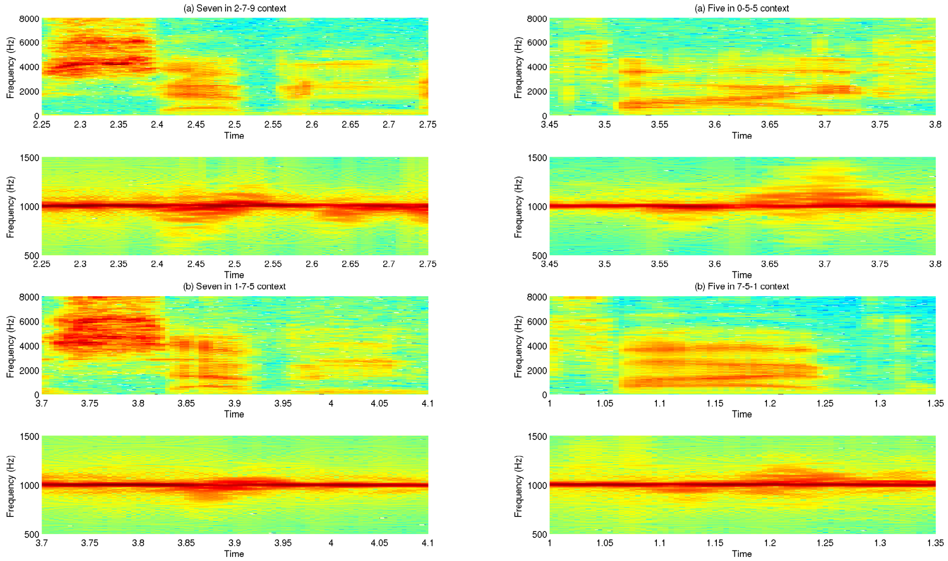


Figure 4-3: Ultrasonic spectrograms in different-context scenarios. The figure setup is the same as in Figure 4-2, but with ultrasonic spectrograms instead of time-domain plots. Much greater similarities can be observed between (a) and (b), (c) and (d).

rier spectrum. For each frame, we then normalize the magnitude of the spectrum by matching its value at the carrier with that of the utterance frame's carrier magnitude. This normalized spectrum is subtracted from each frame of the received spectrum. The bottom spectrogram of Figure 4-4 shows the result of this carrier cancellation. We can observe much more detail in the frequencies near the carrier, which were obscured previously.

4.2.4 Frequency-band energy averages

We have determined from analyzing the ultrasonic signal spectrum that there are consistent trends in the data. There now needs to be a way of quantifying these visible trends for use in machine classification procedures, which require feature vectors as input.

The first type of ultrasonic feature extraction is a simple sub-band averaging method. Figure 4-6 illustrates the spectrum partitioning method. In practice, we partition the ultrasonic spectrum into fourteen non-linearly spaced sub-bands cen-

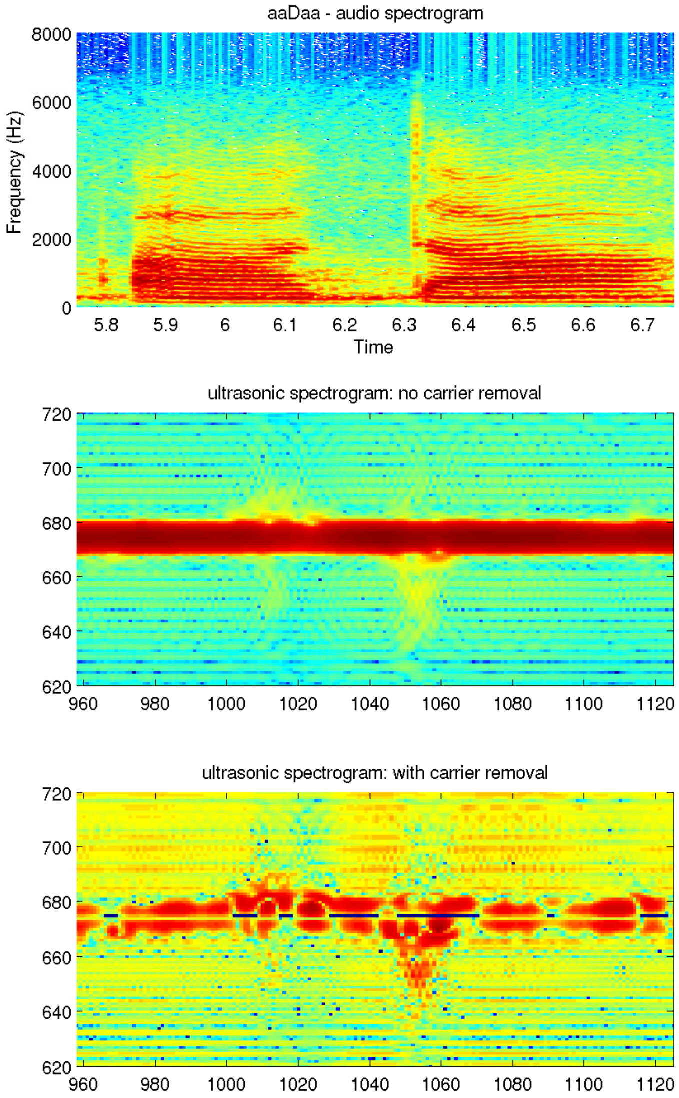


Figure 4-4: Carrier cancellation effects on utterance “aaDaa”. The middle spectrogram shows the carrier coupling signal overwhelming useful received data. The bottom figure shows the spectrogram with the carrier removed.

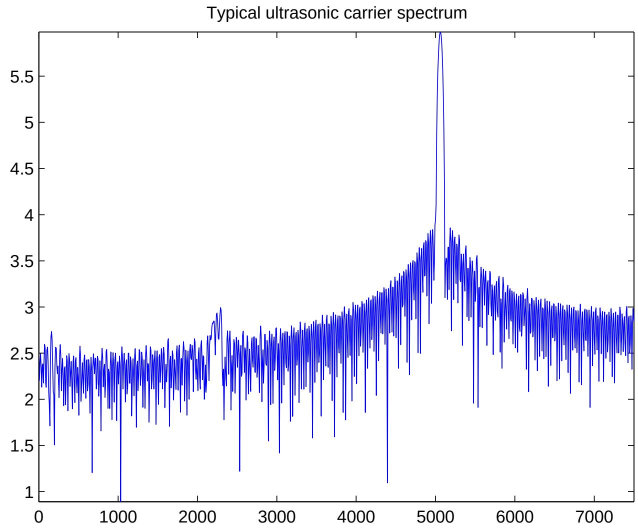


Figure 4-5: A standard spectrum of the ultrasonic carrier signal that will be removed.

tered around the carrier frequency of 4.4kHz. The spectrum of each frame of an utterance is separated into these bands, and the average magnitude of each band is taken as a feature. The bandwidths slowly increase from 40 Hz to 310 Hz from the first to the seventh band, respectively. The bandwidths near the center are smaller in order to capture higher resolution around the carrier frequency. This approach measures the amount of energy (relative to the carrier tone) in different portions of the spectrum. Let FB_i be the frequency band feature for band i for a particular frame, f be frequency, f_{high_i} and f_{low_i} be the frequency boundaries for band i, and $U(f)$ be the magnitude of the ultrasonic spectrum at frequency f.

$$FB_i = \frac{\sum_{f=f_{low_i}}^{f_{high_i}} U(f)}{f_{high_i} - f_{low_i}} \tag{4.1}$$

Figure 4-7 shows sample results of the sub-band feature extraction. Two feature vectors are shown; they have been extracted from the 3.8-4.1 kHz and 4.6-4.9kHz sub-bands (outlined in blue rectangles). We see that at the peaks (both positive and negative) of the spectrogram, there exists high energy in the frequency bands.

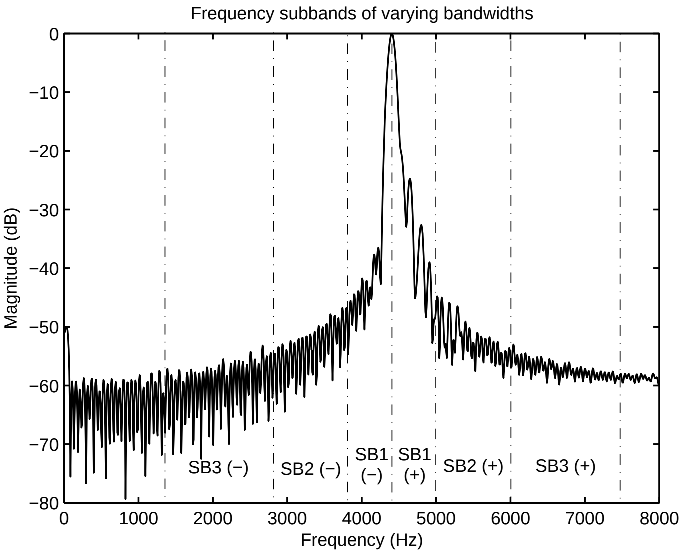


Figure 4-6: Example of six frequency sub-bands on an ultrasonic spectral slice. The average magnitude is computed for each sub-band.

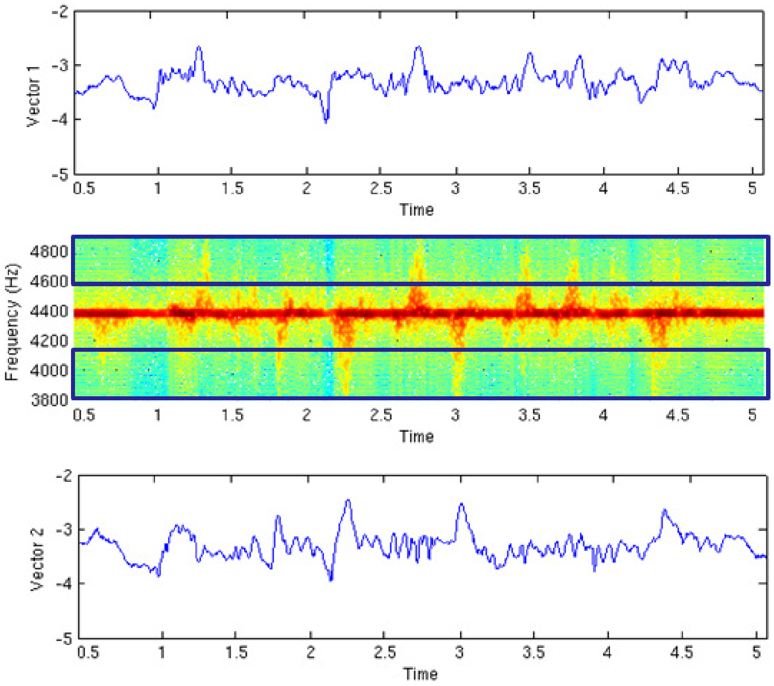


Figure 4-7: Sample frequency sub-band feature vectors obtained from the blue-outlined frequency bands.

4.2.5 Energy-band frequency centroids

The second set of measurements quantifies frequency deviation from the center frequency in different parts of the spectrum. The reasoning for this method of feature extraction is based on the Doppler Effect described in Section 1.2. Frequency deviation from the carrier represents movement in the articulatory surfaces of the user's face. From observation of the ultrasonic spectrogram, we can identify several iso-energy contours, as shown in Figure 4-8. We would like to extract these contours as features.

Figure 4-9 displays the partitioning of the spectrum into several energy bands. The frequencies that exist within each band are weighted by their distance from the carrier frequency, and a center-of-mass (COM) averaging is performed to select one representative frequency centroid for each energy band. Equation 4.2 details the feature extraction calculation. Let EB_j be the energy-band centroid feature for energy band j for a particular frame, f_c be the carrier frequency, and $U'_j(f)$ be a boolean which equals 1 when the frame contains energy in energy band j at frequency f, and 0 when no energy exists at that frequency. Let E^l_j and E^h_j be the low and high energy thresholds for band j, respectively. $U'_j(f)$ acts as a window for a particular energy range, and the EB_j feature is the center-of-mass of the frequency values the window passes through.

$$EB_j = \begin{cases} \dfrac{\displaystyle\sum_{f=f_c}^{8000} \frac{f - f_c}{8000 - f_c} U'_j(f) f}{\displaystyle\sum_{f=f_c}^{8000} \frac{f - f_c}{8000 - f_c} U'_j(f)}, & \text{if } j \text{ band is } > f_c \\ \dfrac{\displaystyle\sum_{f=0}^{f_c} \frac{f_c - f}{f_c} U'_j(f) f}{\displaystyle\sum_{f=0}^{f_c} \frac{f_c - f}{f_c} U'_j(f)}, & \text{if } j \text{ band is } < f_c \end{cases} \tag{4.2}$$

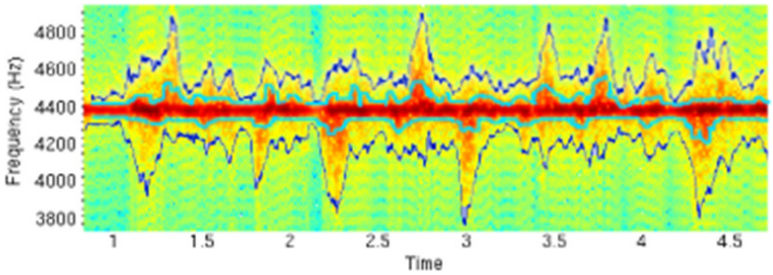


Figure 4-8: Contours on different energy levels.

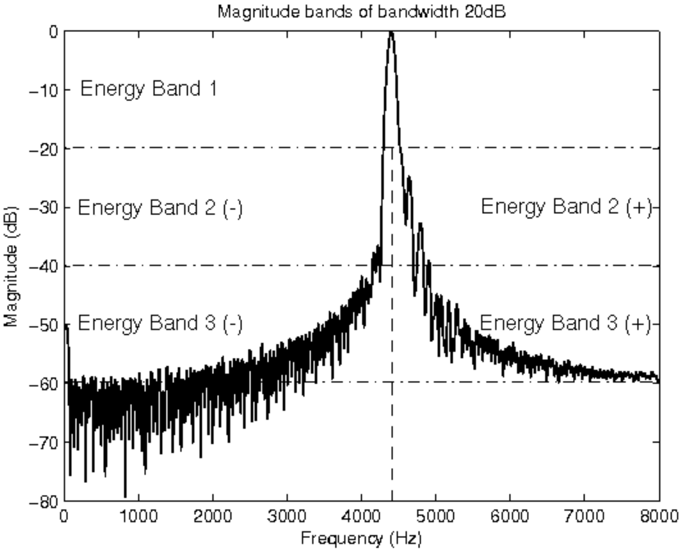


Figure 4-9: Example of five energy sub-bands on an ultrasonic spectral slice. Center-of-mass calculations are performed over frequency ranges defined by relative energy thresholds.

$$U'_j(f) = \begin{cases} 1, & \text{if } U(f) \text{ in } [E^l_j, E^h_j) \\ 0, & \text{otherwise} \end{cases} \tag{4.3}$$

Several energy thresholds were used, over the ranges: -10 to -20 dB, -20 to -30 dB, -30 to -40 dB, -40 to -50 dB, and -50 to -60 dB. Ten total energy band features were computed for each frame.

Figure 4-10 shows sample results of EB feature extraction, from the energy level -50 dB to -60 dB. It is evident that these features closely follow the natural outline of red-to-yellow peaks, which occur at -50 to -60 dB. This outline is directly marked in blue in Figure 4-8.

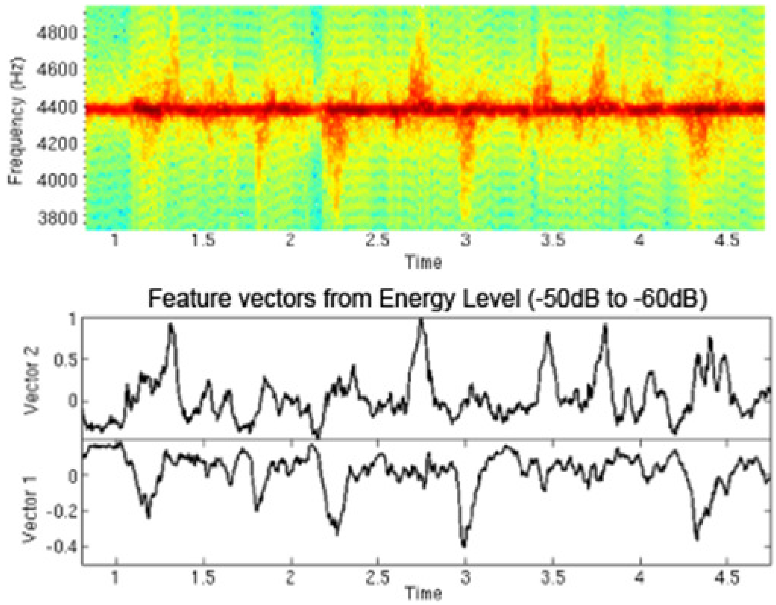


Figure 4-10: Ultrasonic spectrogram of a digit sequence (top), and example feature vectors (bottom).

4.2.6 Peak location features

It is apparent from observing the ultrasonic spectrograms that there are many large peaks in each utterance. These peaks correspond to mouth closures (positive peaks) and mouth openings (negative peaks). Closures cause a high-velocity shifting of the reflection surfaces toward the ultrasonic receiver, thus increasing the observed frequency. Openings cause a high-velocity backwards shift in reflection surfaces, thus decreasing the observed frequency.

Timing information of these closures and openings should provide useful information, especially in relation to phone boundaries. These inter-phone timestamps are calculated through a landmark generation process, which will be detailed in the next section.

Around certain selected landmarks, we find the maximum and minimum peaks of the spectrogram in a 40 ms window (centered at the landmark). To find the peaks, we use two EB features (of the same energy band - one for lower frequencies and and one for upper) as the signal because they can approximate the lower/upper contours along the energy band we are interested in. We are only interested in the larger

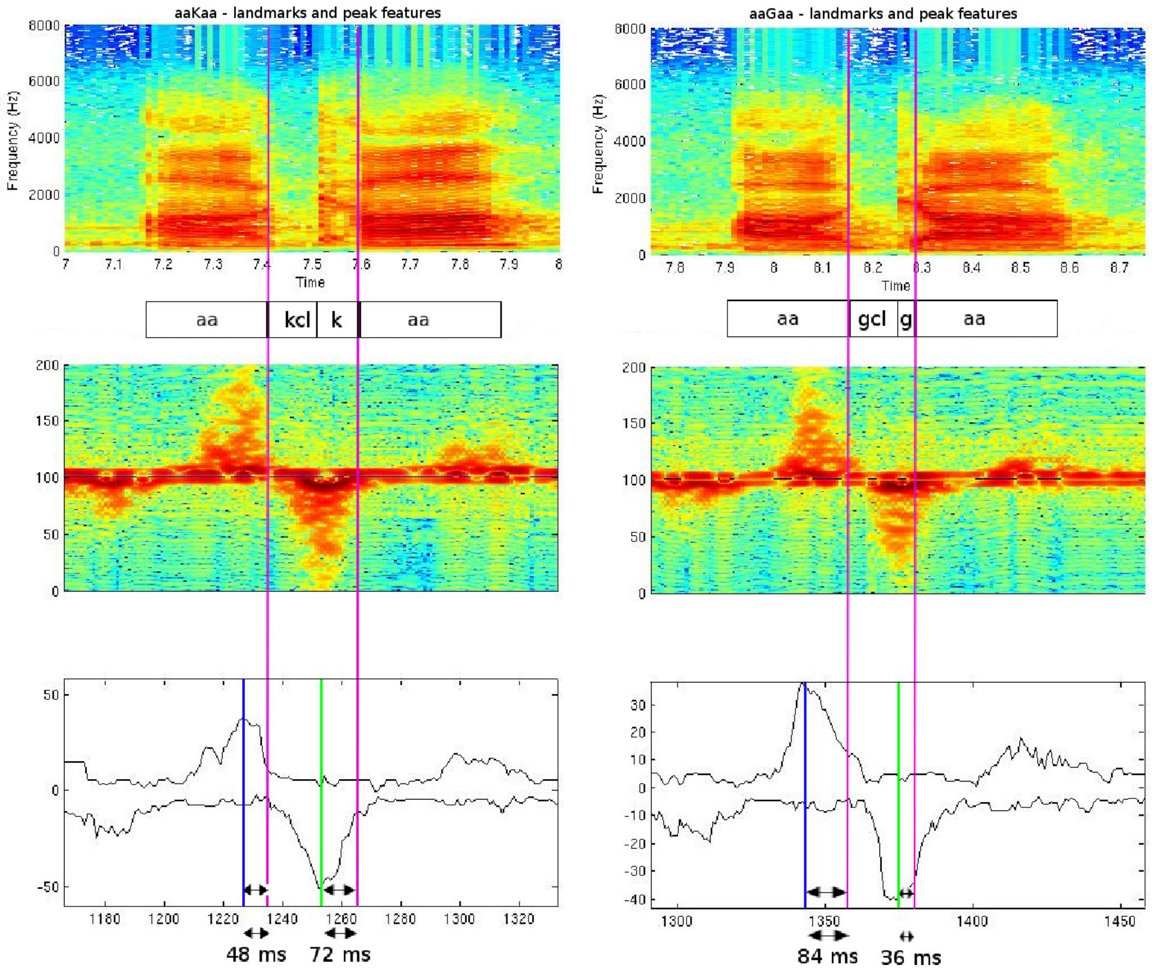


Figure 4-11: Peak location feature extraction process, illustrated with utterances “aaKaa” and “aaGaa”

peak, so we find the difference between the large peak timestamp and the landmark timestamp. Figure 4-11 shows this process visually for the utterances “aaKaa” and “aaGaa”. The landmarks we use are after the first vowel and before the last vowel in each word. These are shown as vertical purple lines which extend through all the sub-figures. The bottom sub-figures show EB extracted contours, with the positive peak marked with blue and the negative peak marked with green. The differences between the peak times and landmark times are also shown; these two difference measurements are the peak location features for that particular word. We observe that these features are different between the two utterances shown.

4.3 Landmark feature processing

Landmarks are used in segment-based speech recognizers [6, 8]. From the acoustic MFCCs, salient changes in acoustic information (hypothetically, phone boundaries) are found and labelled as possible landmarks. Using acoustic models, we then score the segments between all possible landmarks, and then force a one-to-one mapping of the segments to phones (or other acoustic events, such as noise, silence, etc...). This process also automatically selects the "correct" landmarks and rejects the landmarks corresponding to segments with low phonetic likelihood. As explained in Chapter 6, our digit recognizer uses both boundary models, which rely on features around landmark locations, and segment models that are based on the duration of a segment between landmarks.

Using these landmarks, we prepare the audio and ultrasonic features for classification. From streams of frame-based features in each utterance, we end up with an n-dimensional feature vector for each landmark. At each phonetic change landmark, j telescoping windows extend out to each side of the landmark, averaging the features within each window. For k feature sets, there will be a total of $j * k = n$ dimensions for each landmark. Specific dimensionalities will be given in the latter sections regarding recognizer setups.

Chapter 5

Phonetic Classification Experiments

Two types of experiments were performed to investigate the usefulness of ultrasonic information as a second data source for speech recognition. The first set of experiments involved phone and phone-group classification in CVC and VCV contexts, using only the ultrasonic features, to determine the ultrasonic information's ability to distinguish between specific articulatory motions. The second type of experiments involved continuous digit recognition using both audio and ultrasonic features, in which we investigated the effects of varying the ultrasonic model weight as well as acoustic noise levels. We will describe the procedures and findings of the phonetic classification experiments in this section.

5.1 Experimental setup

5.1.1 Data collection procedures

Data collection was performed at MIT in a quiet office environment.

The corpus consisted of eight talkers: six male and two female. The talkers sat in front of the hardware, which captured simultaneous ultrasonic and acoustic data.

The talkers read a script consisting of isolated words each containing a target vowel or consonant. The script consisted of fifteen CVCs in one context ("h-V-d") and twenty-four VCVs in four contexts ("aa-C-aa", "ee-C-ee", "oo-C-oo", "uh-C-uh"), for a total of 111 distinct words. The exact words are in Appendix C. Due to time constraints, we had a different amount of data collected from each individual. Two talkers, one male and one female, contributed 20 sessions (of the entire 111 word collection) of data, while the other talkers contributed 2 sessions each. Thus there are a total of 54 instances of each word. The speaker-dependent experiments were only performed on the two talkers with 20 sessions each.

5.1.2 Classifier setup

As input to the classifier, we generated landmark-based features from only the ultrasonic features for classification. However, in the process of generating the ultrasonic features, we used acoustic information to obtain the landmark locations.

The acoustic data was first run through a recognizer trained on spoken lecture data in forced mode to generate phone boundary landmarks. The correct phoneme sequences were used as input. These landmarks were then edited manually for accuracy.

The edited acoustic landmarks were used to generate ultrasonic features as described in Section 4.2. In a process very similar to that described in Section 6.1.2, the 22 ultrasonic feature streams (10 energy-band frequency centroids and 12 frequency sub-band energy averages) were averaged within twelve telescoping regions around each acoustic landmark (symmetric windows extending 0-6ms, 6-18ms, 18-30ms, 30-60ms, 60-90ms, and 90-180ms on each side from the landmark). Additionally, the two peak features were calculated for each landmark. Each word has two landmarks, placed after the first phone, and before the last phone. When modeling each word, there is a total of 532 dimensions: 528 from the FB and EB features (12 regions * 22 dimensions * 2 landmarks), and 4 from the peak features (2 dimensions * 2 landmarks). Principal components analysis was then used to project down to 50 dimensions, which were modeled with single diagonal Gaussians.

The dimensionality of the models was chosen from a coarse analysis of misclassification rate with respect to dimensionality, and finding the dimensionality for the minimum error rate. The coarse analysis was performed in increments of 10 between 30 and 70 dimensions, and 50 was found to be the optimal dimensionality. Figure 5-1 shows a more recent detailed analysis, with increments of 1, between the ranges of 10 and 70. The classification task was Speaker 2's "aa" context VCVs. We see here that the error rate is minimum at 22 dimensions, although there is also a local minimum at 47 dimensions. Because all the experiments were done with 50-dimension models, classification with 22-dimension models will be future work.

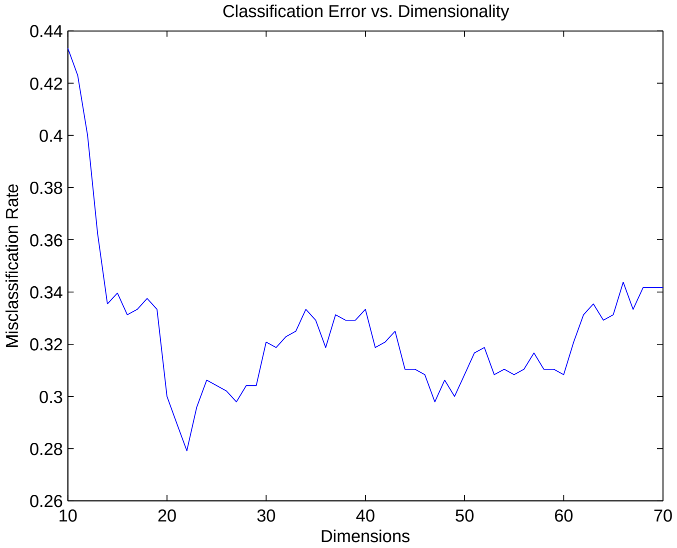


Figure 5-1: Misclassification rate vs. dimensionality.

Several classification experiments were performed, all using the same procedure. We use the jackknife method in obtaining classification results. Given a set of data, 90% is used to train models, and the other 10% is used for testing. Classification performance is measured, and the 90% train/10% test sets are rotated nine more times, resulting in ten sets of classification results, which are then averaged.

We partition the corpus into the contexts denoted above (vowels, "aa", "ee", "oo", and "uh"). Separate classification experiments are performed on each of the consonant contexts to differentiate how they affect our ability to detect consonant articulatory production. We also separately perform speaker-dependent and speaker-independent experiments. In addition to general misclassification measures, we analyze the misclassifications using confusion matrices, which will be discussed in further detail in the next section.

5.2 Preliminary results and observations

For analyzing the classification results, we looked at overall misclassification rates, and investigated in further detail using confusion matrices. We also used raw spectrogram data to understand the trends we observed. In Appendix F, we show one example spectrogram of each word used in this experiment for both Speaker 1 and Speaker 2. Figure 5-2 presents a page of these spectrograms for Speaker 2's "aa" context VCVs.

The analyses we present are preliminary. More data collection and investigation into our feature extraction methods must be done in order to confirm the conclusions. In particular, in deriving the landmark-based features, we may be averaging over rapid changes in the features within the telescoping windows.

5.2.1 Confusion matrix structure

To analyze the classification results, we create a confusion matrix with the two axes representing the hypothesized classes and the correct classes. A sample confusion matrix (Speaker 1's VCV in "aa" context) is shown below in Figure 5-3. The number in each element (A,B) of the matrix indicates the number of times A was classified as B. The shading of each cell is proportional to the classification rate. Darker cells indicate higher classification rates.

In order to simplify the analysis of confusion matrices, we grouped together equivalent misclassifications. For example, a "B" misclassified as "M" is in the same group as "M" misclassified as "B." In both cases, we observe pairwise confusability between "B" and "M". The simplified confusion matrix is shown in Figure 5-4. Notice that the cells in the upper triangle of the matrix have been zeroed out. Those classification rates have been added to their mirror image cells.

Figures D-1 and D-3 in Appendix D show simplified confusion matrices of Speaker 1 and Speaker 2's vowel classification. Figures D-2 and D-4 show confusion matrices

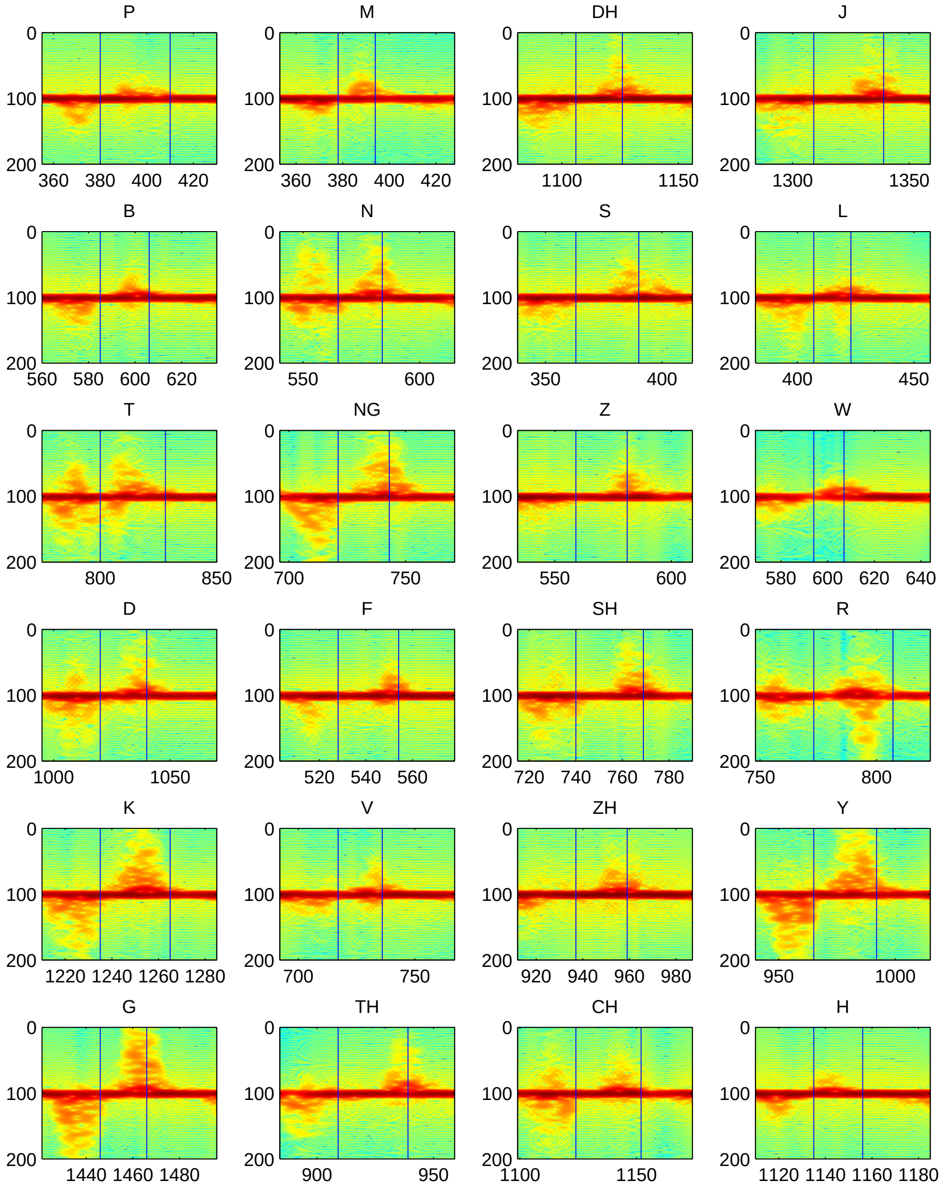
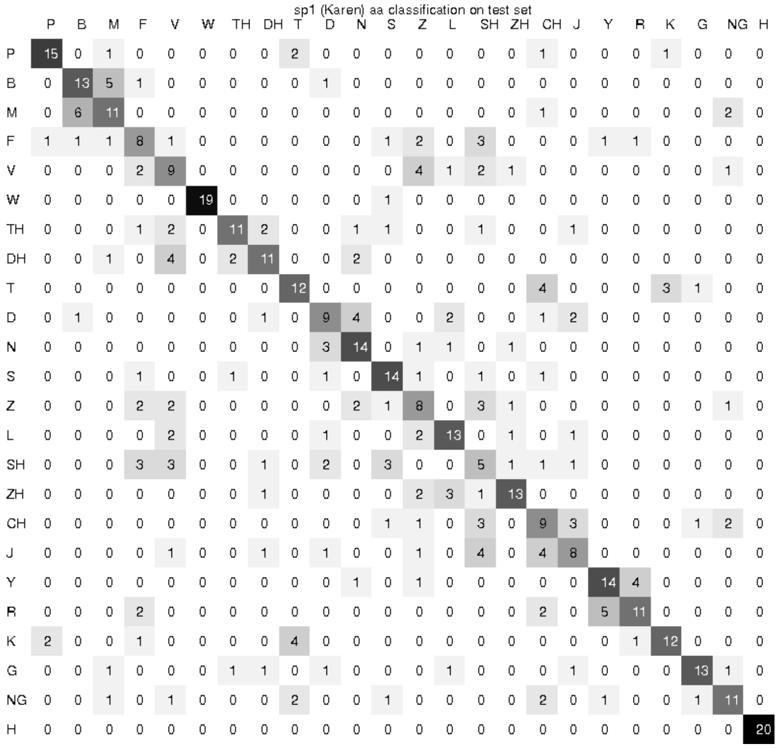


Figure 5-2: Spectrograms of Speaker 2 “aa” context VCVs. Time (in 6ms frames) is represented by the x-axis, while frequency (in 4 Hz frames) is represented by the y-axis.



sp1 (Karen) aa classification on test set

	P	B	M	F	V	W	TH	DH	T	D	N	S	Z	L	SH	ZH	CH	J	Y	R	K	G	NG	H
P	15	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
B	0	13	5	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	6	11	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0
F	1	1	1	8	1	0	0	0	0	0	0	1	2	0	3	0	0	0	1	1	0	0	0	0
V	0	0	0	2	9	0	0	0	0	0	0	0	4	1	2	1	0	0	0	0	0	0	1	0
W	0	0	0	0	0	19	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
TH	0	0	0	1	2	0	11	2	0	0	1	1	0	0	1	0	0	1	0	0	0	0	0	0
DH	0	0	1	0	4	0	2	11	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	4	0	0	0	3	1	0	0
D	0	1	0	0	0	0	0	1	0	9	4	0	0	2	0	0	1	2	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	3	14	0	1	1	0	1	0	0	0	0	0	0	0	0
S	0	0	0	1	0	0	1	0	0	1	0	14	1	0	1	0	1	0	0	0	0	0	0	0
Z	0	0	0	2	2	0	0	0	0	0	2	1	8	0	3	1	0	0	0	0	0	0	1	0
L	0	0	0	0	2	0	0	0	0	1	0	0	2	13	0	1	0	1	0	0	0	0	0	0
SH	0	0	0	3	3	0	0	1	0	2	0	3	0	0	5	1	1	1	0	0	0	0	0	0
ZH	0	0	0	0	0	0	0	1	0	0	0	0	2	3	1	13	0	0	0	0	0	0	0	0
CH	0	0	0	0	0	0	0	0	0	0	0	1	1	0	3	0	9	3	0	0	0	1	2	0
J	0	0	0	0	1	0	0	1	0	1	0	0	1	0	4	0	4	8	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	14	4	0	0	0	0
R	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	5	11	0	0	0	0
K	2	0	0	1	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	1	12	0	0	0
G	0	0	1	0	0	0	1	1	0	1	0	0	0	1	0	0	0	1	0	0	0	13	1	0
NG	0	0	1	0	1	0	0	0	2	0	0	1	0	0	0	0	2	0	1	0	0	1	11	0
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20

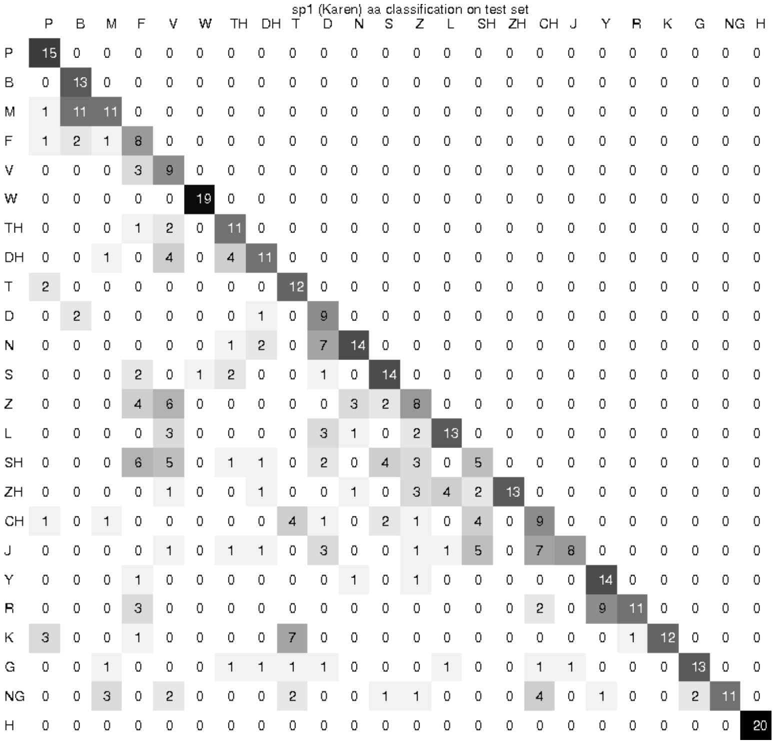
Figure 5-3: Confusion matrix of Speaker 1's "aa" VCV classification.

of Speaker 1 and Speaker 2's VCV classifications that have been superimposed across context. We added the classification results from each of the VCV contexts and compiled them into one confusion matrix.

5.2.2 Acoustic confusability comparisons

We find that certain acoustically confusable pairs such as "M-N" are very rarely confused by the ultrasound-based classifier. On an acoustic spectrogram, the nasals are difficult to distinguish. However, from the confusion matrices in Figures D-2 and D-4, we see that all three pairs "M-N", "M-NG", and "N-NG" have few confusions. From the raw ultrasonic spectrograms in Figure 5-5 we see that the three nasals have different profiles.

In addition to the nasals, we see that other acoustically similar pairs such as "P-B" and "T-D" also have few misclassifications. This low ultrasonic confusability of acoustically confusable phones suggests that even without noise, there are situations in which the ultrasonic signal can contribute usefully, which is important for a



sp1 (Karen) aa classification on test set

	P	B	M	F	V	W	TH	DH	T	D	N	S	Z	L	SH	ZH	CH	J	Y	R	K	G	NG	H
P	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	1	11	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	1	2	1	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	0	0	0	3	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TH	0	0	0	1	2	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DH	0	0	1	0	4	0	4	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	2	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	2	0	0	0	0	0	1	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	1	2	0	7	14	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	2	0	1	2	0	0	1	0	14	0	0	0	0	0	0	0	0	0	0	0	0
Z	0	0	0	4	6	0	0	0	0	0	3	2	8	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	3	0	0	0	0	3	1	0	2	13	0	0	0	0	0	0	0	0	0	0
SH	0	0	0	6	5	0	1	1	0	2	0	4	3	0	5	0	0	0	0	0	0	0	0	0
ZH	0	0	0	0	1	0	0	1	0	0	1	0	3	4	2	13	0	0	0	0	0	0	0	0
CH	1	0	1	0	0	0	0	0	4	1	0	2	1	0	4	0	9	0	0	0	0	0	0	0
J	0	0	0	0	1	0	1	1	0	3	0	0	1	1	5	0	7	8	0	0	0	0	0	0
Y	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	14	0	0	0	0	0
R	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	2	0	9	11	0	0	0	0
K	3	0	0	1	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	1	12	0	0	0
G	0	0	1	0	0	0	1	1	1	1	0	0	0	1	0	0	1	1	0	0	0	13	0	0
NG	0	0	3	0	2	0	0	0	2	0	0	1	1	0	0	0	4	0	1	0	0	2	11	0
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20

Figure 5-4: Simplified confusion matrix of Speaker 1’s “aa” VCV classification.

multimodal system.

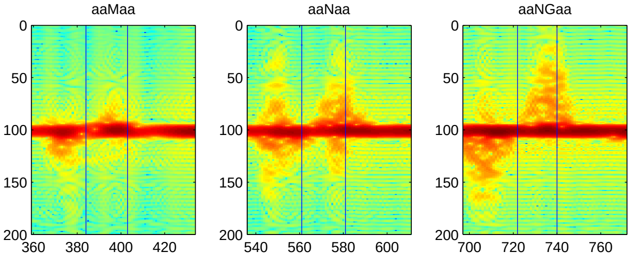


Figure 5-5: Speaker 2’s “aaMaa”, “aaNaa”, and “aaNGaa” spectrograms with acoustic landmarks.

5.2.3 Place of articulation confusability comparisons

We expect that consonant pairs with the same place of articulation would be highly confusable. While this is true for many pairs such as “F-V”, “G-NG”, and “B-M,” we find some surprising exceptions. Along with two examples stated in the previous sub-section, “P-B”(bilabial) and “T-D” (alveolar), we also find “T-N” (alveolar), “K-NG”(velar), and “ZH-CH” (post-alveolar) to be rarely misclassified.

Looking at the spectrograms in Figure 5-6, we see that although “aaKaa” and “aaNGaa” have similar profiles, their peaks occur at different times relative to the vertical landmarks shown (which indicate the end of the first “aa” and the onset of the second “aa”). The velar movement in “aaNGaa” begins long before the first vowel ends, while that of “aaKaa” begins only slightly before the vowel ends. This discrepancy can be seen by the ultrasonic valley of “aaNGaa” occurring before that of “aaKaa”, relative to the first landmark. Similarly, the aspiration after the release of “K” delays the onset of the vowel, while this does not occur with “NG”. Therefore, the second landmark which indicates vowel onset occurs after the second peak of “K”, while the landmark for “NG” occurs during the peak. These discrepancies as well as similar discrepancies for the other rarely misclassified pairs are captured by the features.

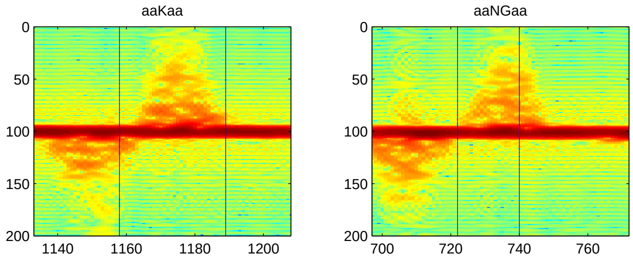


Figure 5-6: Speaker 2’s “aaKaa” and “aaNGaa” spectrograms with acoustic landmarks.

Another expectation we have is that across different places of articulation, we would see few misclassifications. From the confusion matrices in Figures D-2 and D-4, which have the classes arranged on the axes by place of articulation, we see that many of the highly misclassified pairs occur near the diagonal, which means that misclassification often occurs within place of articulation. However, in the areas beyond the diagonal there are numerous highly misclassified pairs. This indicates that there are many phoneme pairs that appear similar in the ultrasonic signal even when they are not in the same place of articulation. An example is “S-TH”, whose similarities are demonstrated by the spectrograms in Figure 5-7.

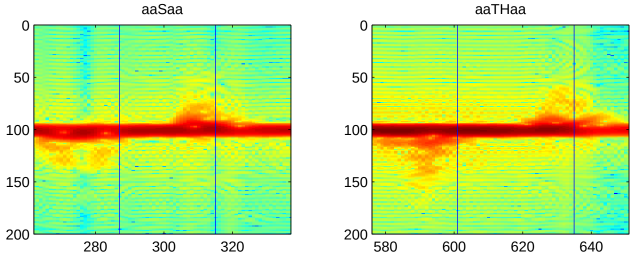


Figure 5-7: Speaker 2's "aaSaa" and "aaTHaa" spectrograms with acoustic landmarks.

5.2.4 Context and speaker dependency

We have performed consonant classification tasks on four separate vowel contexts ("aa", "ee", "oo", and "uh"). Table 5.1 shows speaker-dependent misclassification measurements across these contexts. We see here that classification widely varies depending on the context.

Within consonant classification, we observe that consonants in "ee/oo" contexts are more difficult to classify than those in "aa/uh" contexts. This may be due to the more closed mouth positions of "ee/oo" which result in smaller articulatory movements than in "aa/uh" contexts. From ultrasonic spectrograms, we can see evidence for this in the lower frequency deviation from the carrier frequency in the "ee/oo" contexts. Figure 5-8 shows "vowel-NG-vowel" across the four contexts.

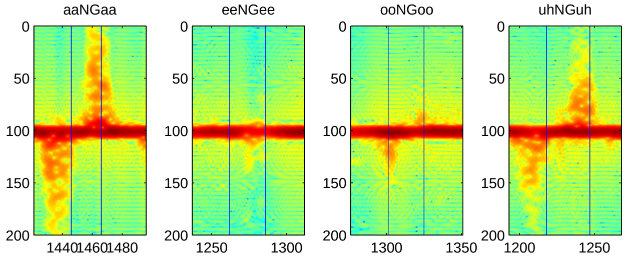


Figure 5-8: Speaker 2's "aaNGaa", "eeNGee", "ooNGoo", and "uhNGuh" spectrograms with acoustic landmarks.

With less mouth movement, there is less ultrasonic signal resolution in the features and class models. The most prominent information comes from the openings and closings of the lips because those movements effectively move a large area of reflecting surfaces toward/away from the receiver very quickly. This almost-instantaneous shift

Speaker 1 (F) Misclassification (%)		Speaker 2 (M) Misclassification (%)	
Context	Test	Context	Test
vowel	60.00	vowel	50.33
aaCaa	41.04	aaCaa	30.83
eeCee	66.25	eeCee	52.08
ooCoo	71.25	ooCoo	45.21
uhCuh	44.17	uhCuh	41.67

Table 5.1: Speaker 1 (Female) and Speaker 2 (Male) Speaker-Dependent Misclassification Rates

Speaker Independent) Misclassification (%)	
Context	Test
vowel	75.47
aaCaa	57.08
eeCee	74.92
ooCoo	78.58
uhCuh	65.08

Table 5.2: Speaker-Independent Misclassification Rates

causes large frequency shifts in the ultrasonic signal. Additionally, with a more open context, we are also able to observe tongue movements better.

In Appendix D, Figures D-5 through D-12 show specific confusion matrices of VCVs in the four contexts for both speakers. From glancing over these matrices it is evident that there are many differences in misclassifications across contexts. Appendix E contains tables of the 10 most misclassified pairs for each context; we see from Tables E.1 and E.2 that highly confusable pairs differ across the vowel contexts.

From the same few figures and tables mentioned above, we observe differences between Speaker 1 and Speaker 2's misclassification results. Speaker 2's models generally perform better, and given a context, the highly confusable pairs differ amongst the speakers, as shown by Tables E.1 and E.2. It is no surprise then that speaker-dependent classification performance (Table 5.1) is much better than that of speaker-independent classification (Table 5.2).

5.2.5 Hierarchical clustering analysis

For the "aa" context, Figures 5-9 and 5-10 show dendrograms of hierarchical clustering analysis of the consonant classes. The distance between two classes was computed by subtracting the misclassification rate for that class pair from the maximum misclassification rate over all pairs in the experiment. Thus, higher error rates correspond to smaller distances between classes. The dendrograms were created with the shortest distance method in MATLAB.

From these dendrograms we can understand how close two phonemes are to each other with respect to classification rates. For both speakers, the "B-M" pair was highly confusable, as evidenced by the confusion matrices in Figures D-5 and D-9. This is reflected by the clustering of "B" and "M" in the lowest level of the dendrograms. Similarly, "W" and "H" were rarely confused with any other phones, so they were clustered last. In the middle cases, other inter-speaker similarities as well as dissimilarities can be observed from these dendrograms.

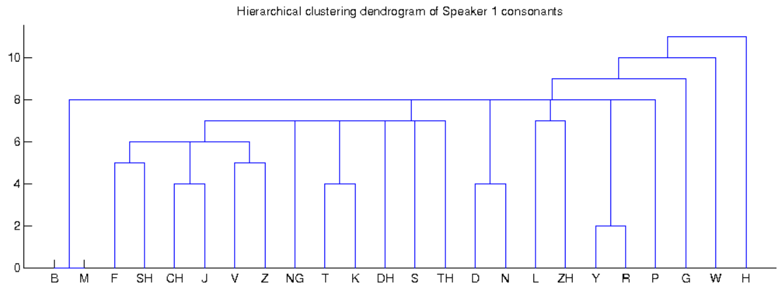


Figure 5-9: Speaker 1's dendrogram for "aa" context.

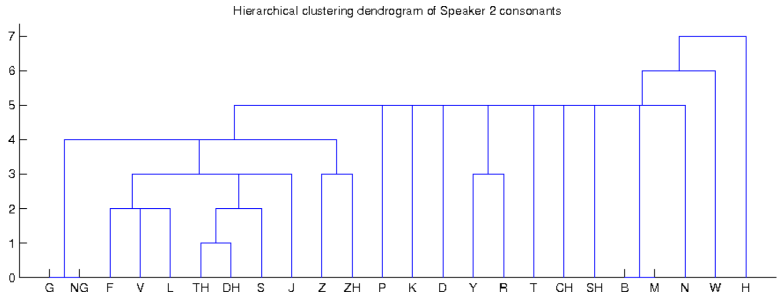


Figure 5-10: Speaker 2's dendrogram for "aa" context.

Chapter 6

Digit Recognition Experiments

For this experiment, we performed continuous digit recognition, while varying the weight given to the ultrasonic model as well as the level of acoustic input noise.

6.1 Experimental setup

6.1.1 Data collection procedures

Data collection was performed at MIT in a quiet office environment.

The corpus consisted of twenty talkers: nineteen male and one female. The talkers were situated in front of the ultrasonic transducers, with a distance of about six inches between the talker's face and the transducers. The talkers were told to limit their head movement as much as possible. The microphone on the hardware simultaneously captured acoustic data.

The talkers were prompted with fifty sequences, each containing ten randomized digits. These sequences can be referenced in Appendix B. The digits were 0 through 9, and the users were told to say "zero" instead of "oh" for consistency. The entire data set consisted of one thousand ten-digit utterances; each digit was spoken approximately one thousand times. For our experiments, we divided our collected data into a training set containing 750 utterances from 15 speakers, and a test set containing 250 utterances from the remaining set of 5 speakers.

6.1.2 Recognizer setup

Our speech recognition experiments were conducted using a landmark-based speech recognizer that has been previously used for AVSR experiments [6, 8]. The recognizer was configured to recognize arbitrary digit strings containing exactly 10 digits. The digit strings were modeled by 110 context-dependent diphone-based acoustic and ultrasonic models.

To generate the landmark-based acoustic features, the speech signal was first processed into frame-based Mel-frequency scale cepstral coefficients (MFCCs) at a rate of 200 frames per second. Each frame consisted of a vector of 14 MFCCs, which were described in Section 4.1. From the MFCC frames, significant landmarks in the acoustic signal were first detected using a measure of acoustic change. Feature vectors were extracted at landmarks based on averages of MFCC vectors in the region surrounding each landmark. Specifically, a set of 8 telescoping regions were defined, which together span 150ms around the landmark (symmetric windows extending 0-5ms, 5-15 ms, 15-35ms, and 35-75ms on each side of the landmark). Within each of these regions the frame-based MFCC feature vectors were averaged to form a single 14-dimensional feature vector for the entire region. In total, this yielded a single 112-dimensional (8 regions * 14 dimensions) feature vector for each landmark. The landmark feature vectors were then projected down to 50 dimensions using principal components analysis (PCA). From the 50-dimensional feature vectors extracted from the training data, word-dependent diphone-based phonetic models were created to represent the acoustic landmarks within the digit words. Gaussian mixture density functions were used to model the 110 diphone models.

The models of the ultrasonic features were generated in a similar fashion as the acoustic models. For every frame the ultrasonic signal was represented by the collection of 27 ultrasonic measurements (13 energy-band frequency centroids and 14 frequency sub-band energy averages). Within each of six telescoping regions surrounding an acoustic landmark, the ultrasonic frame vectors were averaged to form a single 27-dimension feature vector for the entire region. The full set of six regions spans 140ms around the landmark (symmetric windows extending 0-10ms, 10-30ms, and 30-70ms out each side of the landmark). In total, this yields a 162-dimensional (6 regions * 27 dimensions) ultrasonic feature vector for each landmark. The ultrasonic landmark feature vectors were then also projected down to 35 dimensions using principal components analysis. As with the acoustic features, the ultrasonic features were modeled with a Gaussian mixture density function for each of the 110 different diphone models.

In addition to the acoustic and ultrasonic models, a context independent phonetic duration model was also created [25]. The three models were trained on the data in the 15 speaker training set. In the baseline recognizer configuration, the acoustic, ultrasonic and duration models were combined with equal weights of 1. In situations where there may be considerable background acoustic noise, the system can reduce the weight of the acoustic model relative to the ultrasonic model as the acoustic signal-to-noise ratio (SNR) is reduced.

To simulate noisy acoustic conditions, babble noise from the NOISEX database was synthetically added to the data in the test set at SNR levels of 20db, 10db and 0db [23]. This provided us with four noise conditions (including the clean condition) for our experiments. At each noise condition we examined the recognition performance as the weight of the acoustic model was varied from 0.0 to 1.0.

6.2 Results and Discussion

In general, we have found that using ultrasonic information in addition to acoustic information improves digit recognition performance.

In Figure 6-1, we see results from all four acoustic noise-level settings. The x-axis represents audio weight, while the y-axis shows Word Error Rate (WER) on a logarithmic scale. A solid curve for each noise condition shows the multimodal (audio+ultrasonic) recognition results as the audio weight is varied from 0.0 to 1.0, and a dashed line is shown for the unimodal audio-only result. The graph shows that the ultrasonic information improves the speech recognition performance over

Noise Level	Optimal Audio Weight	Word Error Rate (%) Audio Only	Ultrasonic Only	Audio + Ultrasonic
Clean	1.0	0.32	70.5	0.24
20db	1.0	3.44	70.5	2.44
10db	0.5	24.0	71.8	17.5
0db	0.3	61.2	72.0	46.6

Table 6.1: Digit recognition results for the audio-only, ultrasonic-only, and multi-modal (audio+ultrasonic) systems when the optimal audio weight is used.

the audio-only case for a wide range of audio weights for each condition. This is confirmation that ultrasonic data is a useful secondary modality for noise-robust speech recognition.

For each noise level, there is an optimal audio weighting which provides the best recognition result, i.e. the minimum WER. These optimal points are circled (in red) on the chart. An important point to note is that as the noise level increases, the optimal audio weight decreases (1.0 to 0.5 to 0.3). This demonstrates that with increasing noise, the audio information becomes less important, and the ultrasonic data contributes more to accurate recognition. This is expected because ultrasonic data should be immune to acoustic noise, and its usefulness should increase relative to acoustic data with added acoustic noise.

Even without optimal audio weighting, i.e. keeping the audio weight at a baseline level of 1.0, we see by the green boxes on the chart that we still obtain similar improvements over the audio-only scenario.

Table 6.1 summarizes the results from the figure. The audio+ultrasonic performance is presented at the optimal audio weight setting. Over the four different noise conditions, relative error rate reductions from audio-only to the audio+ultrasonic system varied between 24% and 29%. Notice that the ultrasonic-only performance is quite poor, at around 71% WER (this measurement changes with noise level only because the ultrasonic features depend upon acoustic landmarks, which shift slightly with noisy data). However, the fusion with acoustic data improves the performance significantly.

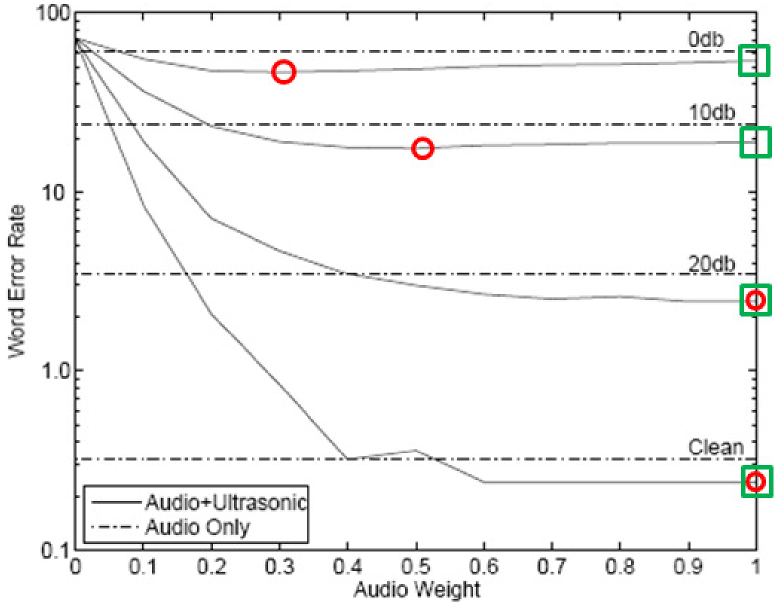


Figure 6-1: Digit recognition results for four noise levels as the audio weight is varied from 0.0 to 1.0. Audio+ultrasonic results are represented by the solid lines, while audio-only results are given by the dashed lines.

Chapter 7

Conclusions

7.1 Summary

In this research we built a multimodal speech recognition system that uses ultrasonic sensing of articulatory movement as a second modality beyond the standard acoustic information. We tested our system on a continuous digit recognition task as well as phoneme and phoneme cluster classification tasks. Our digit recognition experiment demonstrates improved word error rate (WER) performance across multiple noise levels when including ultrasonic data as a second recognition modality.

7.1.1 System description

We built hardware to simultaneously capture acoustic and ultrasonic speech data. In addition to an onboard mic, an ultrasonic transmitter/receiver pair is aimed at the talker's mouth. The transmitter emits a continuous 40 kHz sinusoid, which is reflected by the talker's moving articulators during speech. These movements cause Doppler frequency shifts in the received signal; the frequency shifts are characterized by features we have designed, which are modeled with Gaussian densities.

Three types of features are extracted from the ultrasonic data at each time frame. The first type is the average energy of the signal for a given frequency band of the spectrum. The second type is an averaged frequency deviation (from the carrier) for a given energy band of the spectrum. This feature corresponds to contours running along a certain energy band in the ultrasonic spectrogram. The third feature type, which was only used in the phonetic classification experiments, represents the timing of the mouth closure and opening relative to the beginning and end of the phone.

The features at each frame are further processed into feature vectors for diphone (for digit recognition) or word (for phonetic classification) Gaussian models. Acoustic landmarks are computed, defining the phone boundaries around which these frame features are averaged and concatenated into class feature vectors. The averaging is done over telescoping time windows extending from these landmarks. For the multimodal digit recognition task, acoustic MFCC-based models are also computed. The phonetic classification experiments use only ultrasonic models.

7.1.2 System testing and performance

Phonetic classification

In a preliminary study, we investigated the ultrasound's abilities to distinguish phones in different contexts. More data collection and more precise features appropriate for this task could help to confirm our observed trends. We measured overall misclassification rates as well as analyzed in detail classification confusion matrices.

We have observed that phones that are acoustically similar (such as "m" and "n") are often distinct in the ultrasonic signal because the articulatory motions are different. This provides some evidence for the orthogonality the two sources, which is desirable for a multimodal system. We have seen that the expected confusability between two consonants with the same place of articulation (such as "p" and "b") is often nonexistent. Much of this can be explained by relative timing differences between articulatory events in these phonemes.

Our experiments on consonants were context-dependent. We have seen that different vowel contexts ("aa", "ee", "oo", and "uh") result in different classification results and different confusion matrices. Finally, we have observed that speaker-dependent classification outperformed speaker-independent classification, as expected. The differences between talkers' articulatory styles resulted in dissimilar features that were averaged together in the consequently poor speaker-independent models.

These trends can also be observed qualitatively by comparing pairs of the raw spectrograms visually.

Digit recognition

We performed a continuous speaker-independent digit recognition task, while varying the audio/ultrasonic model weight ratio as well as varying the amount of acoustic noise. Over four noise levels (clean, 20 dB, 10 dB, and 0 dB), the recognizer reduced word error rates by a relative 24% to 29%. At each noise level, there was an optimal audio model weighting which resulted in the best performance. As we increased the noise level, this optimal audio weight decreased, indicating that the ultrasonic information contributes more toward accurate recognition as the audio becomes noisier. The digit recognition experiment demonstrates that ultrasonic information is an effective modality for noise-robust speech recognition.

7.2 Future work

For the phonetic classification experiments, the dimensionality of our features was greatly reduced to 50-dimensions because of data sparsity. This could be solved with more data collection. More users and more data per user would improve the speaker-dependent modeling. More work should be done in feature extraction as well. The current landmark-based method could be averaging features that change quickly over time. By capturing frame-by-frame spectral information, dynamic time warping (DTW), for phonetic classification could be useful; a template for each CVC or VCV word would be warped against. We have already seen similarities and differences across word spectrograms through subjective visual evaluation. Another possible improvement to the phonetic classification task is further analysis of the composition of the automatically generated clusters, and investigating the reason behind certain phones being clustered together.

For recognition tasks, larger vocabulary experiments could be done, beyond the digit domain. Medium vocabulary recognizers for information kiosks could be built and tested. Data collection would then become more automatic, although there would be problems with unsupervised head movements and incorrect facial positioning. With a larger vocabulary, better features for continuous speech would need to be developed.

Beyond speech recognition, there are applications such as speaker verification and gait/walker identification [12] that could be explored further. These systems could be integrated into an existing speech recognition system for purposes such as speaker adaptation or automatic selection of specific speaker-dependent models. Physically, hardware issues could be explored, such as the use of ultrasonic beamforming arrays or placing ultrasonic transducers on a headset for portability.

Appendix A

Hardware Schematics

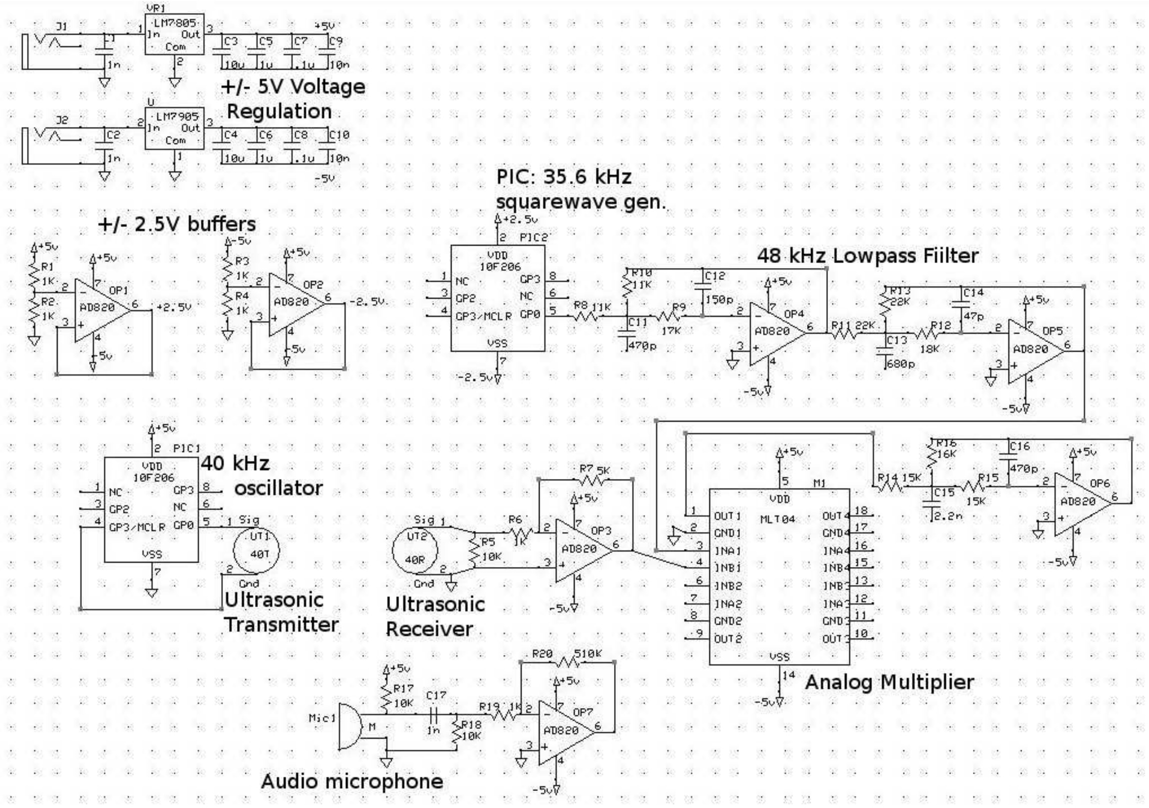


Figure A-1: Annotated schematic of prototype hardware.

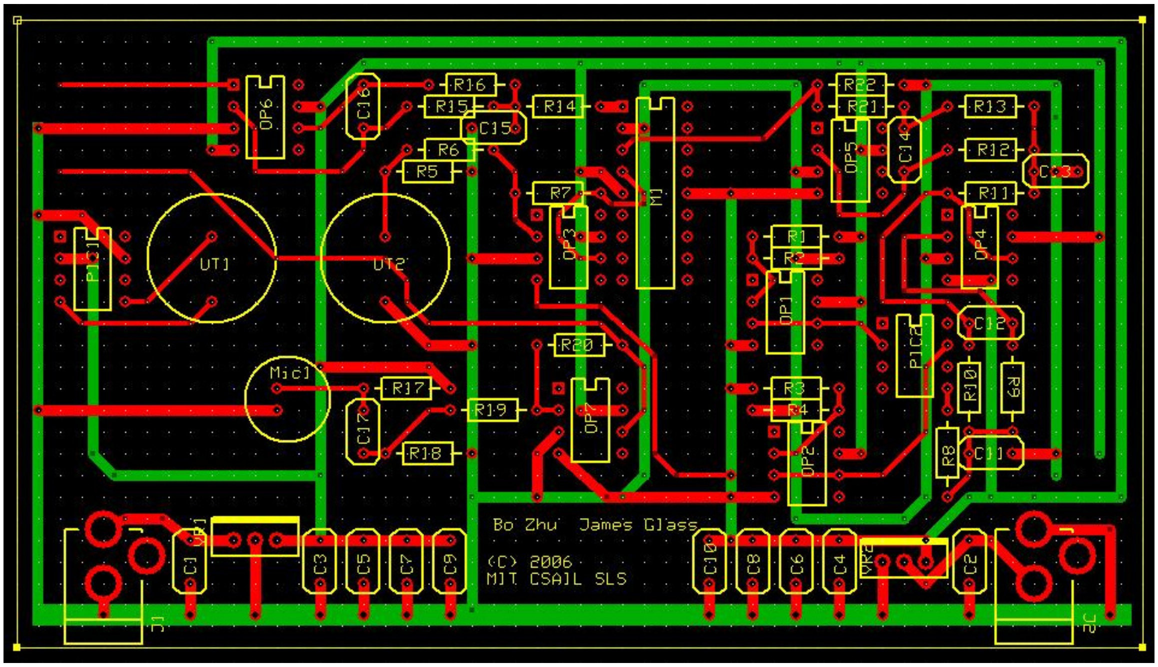


Figure A-2: PCB diagram of hardware layout.

Appendix B

Digit Recognition Utterances

4 6 7 0 5 0 5 5 6 0

7 5 1 9 0 8 4 6 4 9

1 1 7 2 2 1 3 8 6 9

7 0 8 2 7 9 9 1 2 9

2 8 0 8 8 9 0 1 7 5

4 0 3 8 8 6 9 9 2 5

4 3 6 0 4 3 7 6 0 2

8 9 3 2 6 7 5 3 6 9

6 7 0 2 8 0 6 8 9 4

5 4 9 8 1 7 6 4 1 7

3 0 2 2 7 3 2 4 0 9

9 3 2 4 7 6 5 4 9 4

9 4 3 3 6 8 6 2 2 0

4 2 6 7 8 6 8 9 3 1

0 7 1 7 6 7 7 6 0 4

2 9 2 8 0 3 0 1 8 8

2 4 4 5 4 8 5 9 6 8

2 2 2 9 8 4 2 2 3 6

8 8 9 5 2 0 3 1 5 8

5 2 6 2 9 3 8 2 6 3

7 7 9 3 7 7 1 9 5 4

5 4 3 4 4 8 2 2 5 4

0 9 4 8 6 7 6 6 6 3

8 1 1 3 7 8 3 8 4 3

5 0 7 7 5 1 4 5 3 8

7 2 5 9 5 9 3 2 4 9

1 8 7 6 7 6 0 2 5 0

5 2 4 9 5 6 8 4 6 0

4 2 7 5 3 9 3 0 3 0

3 3 3 0 7 5 8 3 5 1

8 8 4 9 9 8 9 9 6 7

4 6 4 4 5 1 8 2 0 5

2 7 5 3 9 7 3 9 9 0

2 7 2 4 0 5 9 6 9 6

1 6 1 4 4 9 8 4 7 2

0 2 8 6 3 3 2 1 1 5

6 9 6 7 8 8 3 4 4 8

2 3 8 8 1 5 4 5 7 9

3 0 2 4 9 8 5 6 1 5

8 3 7 6 8 3 2 8 2 9

1 1 6 1 0 8 0 6 2 2

6 9 9 7 3 8 7 3 4 1

3 8 1 8 9 1 6 6 4 3

0 0 7 3 6 0 7 1 9 1

1 8 9 2 0 6 5 5 6 5

2 1 8 4 0 8 7 8 7 0

0 3 4 7 3 4 3 9 7 8

5 6 6 7 4 6 1 5 7 4

4 1 4 8 7 7 1 9 3 4

1 6 1 5 9 7 8 4 7 4

Appendix C

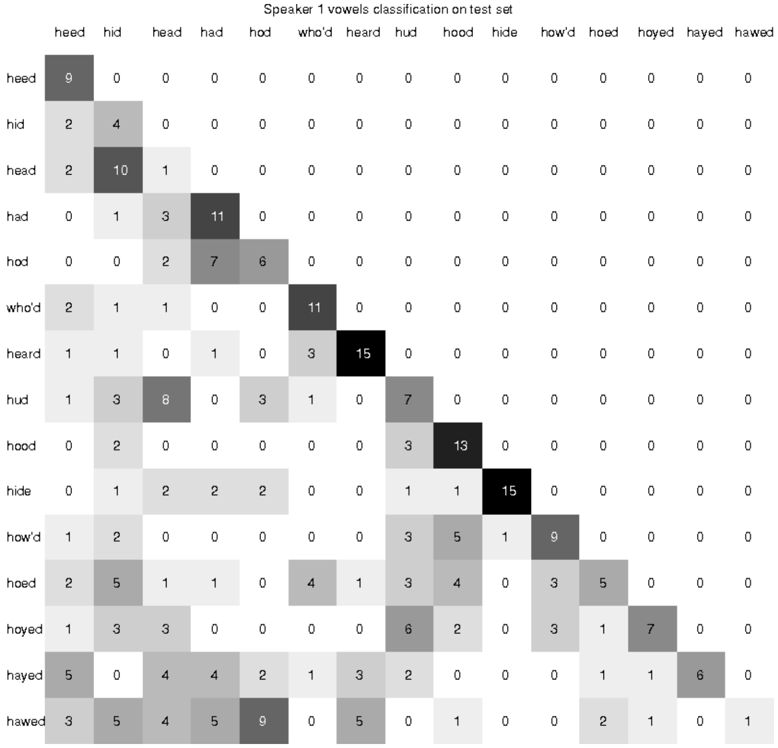
Phonetic Classification Utterances

heed	aaPaa	ePee	ooPoo	uh-Puh
hid	aaBaa	eeBee	ooBoo	uh-Buh
head	aaTaa	eeTee	ooToo	uh-Tuh
had	aaDaa	eeDee	ooDoo	uh-Duh
hod	aaKaa	eeKee	ooKoo	uh-Kuh
who'd	aaGaa	eeGee	ooGoo	uh-Guh
heard	aaMaa	eeMee	ooMoo	uh-Muh
hud	aaNaa	eeNee	ooNoo	uh-Nuh
hood	aaNGaa	eeNGee	ooNGoo	uh-NGuh
hide	aaFaa	eeFee	ooFoo	uh-Fuh
how'd	aaVaa	eeVee	ooVoo	uh-Vuh
hoed	aaTHaa	eeTHee	ooTHoo	uh-THuh
hoyed	aaDHaa	eeDHee	ooDHoo	uh-DHuh
hayed	aaSaa	eeSee	ooSoo	uh-Suh
hawed	aaZaa	eeZee	ooZoo	uh-Zuh
	aaSHaa	eeSHee	ooSHoo	uh-SHuh
	aaZHaa	eeZHee	ooZHoo	uh-ZHuh
	aaCHaa	eeCHee	ooCHoo	uh-CHuh
	aaJaa	eeJee	ooJoo	uh-Juh
	aaLaa	eeLee	ooLoo	uh-Luh
	aaWaa	eeWee	ooWoo	uh-Wuh
	aaRaa	eeRee	ooRoo	uh-Ruh
	aaYaa	eeYee	ooYoo	uh-Yuh
	aaHaa	eeHee	ooHoo	uh-Huh

Table C.1: Phonetic Classification Utterances. The capitalized target consonants are represented here using the ARPAbet phonetic alphabet. The speakers were instructed in their proper pronunciation.

Appendix D

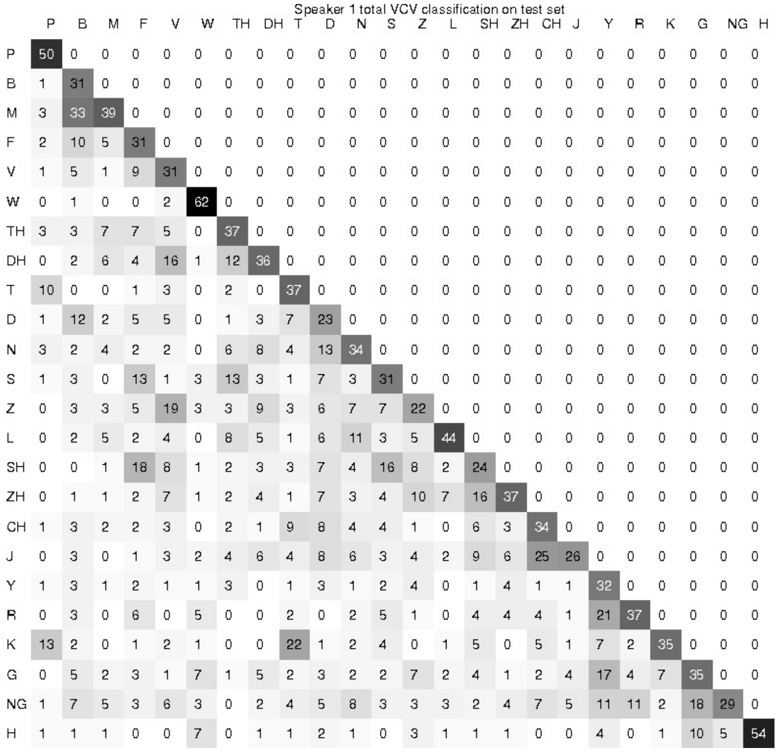
Referenced Confusion Matrices



Speaker 1 vowels classification on test set

	heed	hid	head	had	hod	who'd	heard	hud	hood	hide	how'd	hoed	hoyed	hayed	hawed
heed	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hid	2	4	0	0	0	0	0	0	0	0	0	0	0	0	0
head	2	10	1	0	0	0	0	0	0	0	0	0	0	0	0
had	0	1	3	11	0	0	0	0	0	0	0	0	0	0	0
hod	0	0	2	7	6	0	0	0	0	0	0	0	0	0	0
who'd	2	1	1	0	0	11	0	0	0	0	0	0	0	0	0
heard	1	1	0	1	0	3	15	0	0	0	0	0	0	0	0
hud	1	3	8	0	3	1	0	7	0	0	0	0	0	0	0
hood	0	2	0	0	0	0	0	3	13	0	0	0	0	0	0
hide	0	1	2	2	2	0	0	1	1	15	0	0	0	0	0
how'd	1	2	0	0	0	0	0	3	5	1	9	0	0	0	0
hoed	2	5	1	1	0	4	1	3	4	0	3	5	0	0	0
hoyed	1	3	3	0	0	0	0	6	2	0	3	1	7	0	0
hayed	5	0	4	4	2	1	3	2	0	0	0	1	1	6	0
hawed	3	5	4	5	9	0	5	0	1	0	0	2	1	0	1

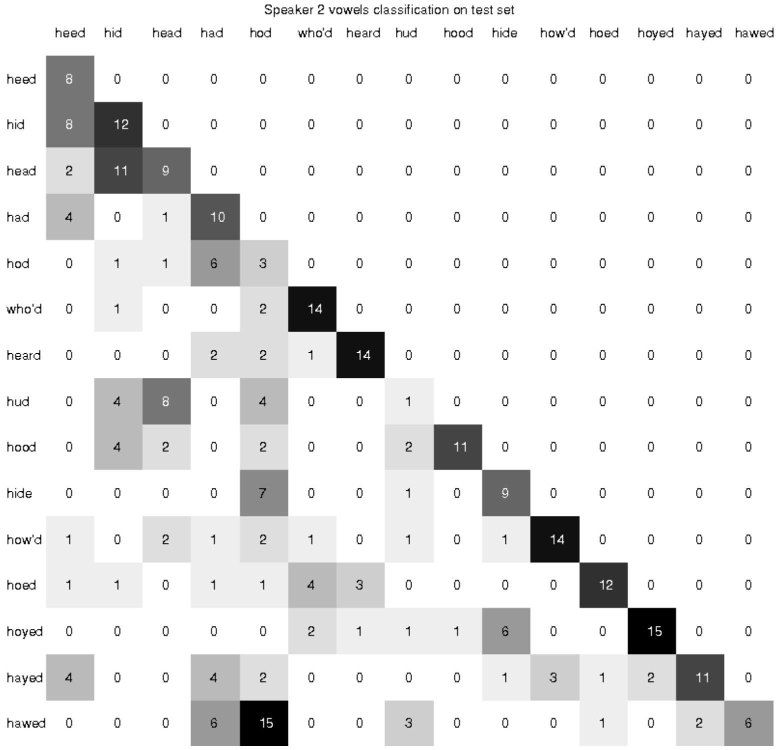
Figure D-1: Simplified confusion matrix of Speaker 1's vowel classification.



Speaker 1 total VCV classification on test set

	P	B	M	F	V	W	TH	DH	T	D	N	S	Z	L	SH	ZH	CH	J	Y	R	K	G	NG	H
P	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	1	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	3	33	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	2	10	5	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	1	5	1	9	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	1	0	0	2	62	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TH	3	3	7	7	5	0	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DH	0	2	6	4	16	1	12	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	10	0	0	1	3	0	2	0	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	1	12	2	5	5	0	1	3	7	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	3	2	4	2	2	0	6	8	4	13	34	0	0	0	0	0	0	0	0	0	0	0	0	0
S	1	3	0	13	1	3	13	3	1	7	3	31	0	0	0	0	0	0	0	0	0	0	0	0
Z	0	3	3	5	19	3	3	9	3	6	7	7	22	0	0	0	0	0	0	0	0	0	0	0
L	0	2	5	2	4	0	8	5	1	6	11	3	5	44	0	0	0	0	0	0	0	0	0	0
SH	0	0	1	18	8	1	2	3	3	7	4	16	8	2	24	0	0	0	0	0	0	0	0	0
ZH	0	1	1	2	7	1	2	4	1	7	3	4	10	7	16	37	0	0	0	0	0	0	0	0
CH	1	3	2	2	3	0	2	1	9	8	4	4	1	0	6	3	34	0	0	0	0	0	0	0
J	0	3	0	1	3	2	4	6	4	8	6	3	4	2	9	6	25	26	0	0	0	0	0	0
Y	1	3	1	2	1	1	3	0	1	3	1	2	4	0	1	4	1	1	32	0	0	0	0	0
R	0	3	0	6	0	5	0	0	2	0	2	5	1	0	4	4	4	1	21	37	0	0	0	0
K	13	2	0	1	2	1	0	0	22	1	2	4	0	1	5	0	5	1	7	2	35	0	0	0
G	0	5	2	3	1	7	1	5	2	3	2	2	7	2	4	1	2	4	17	4	7	35	0	0
NG	1	7	5	3	6	3	0	2	4	5	8	3	3	3	2	4	7	5	11	11	2	18	29	0
H	1	1	1	0	0	7	0	1	1	2	1	0	3	1	1	1	0	0	4	0	1	10	5	54

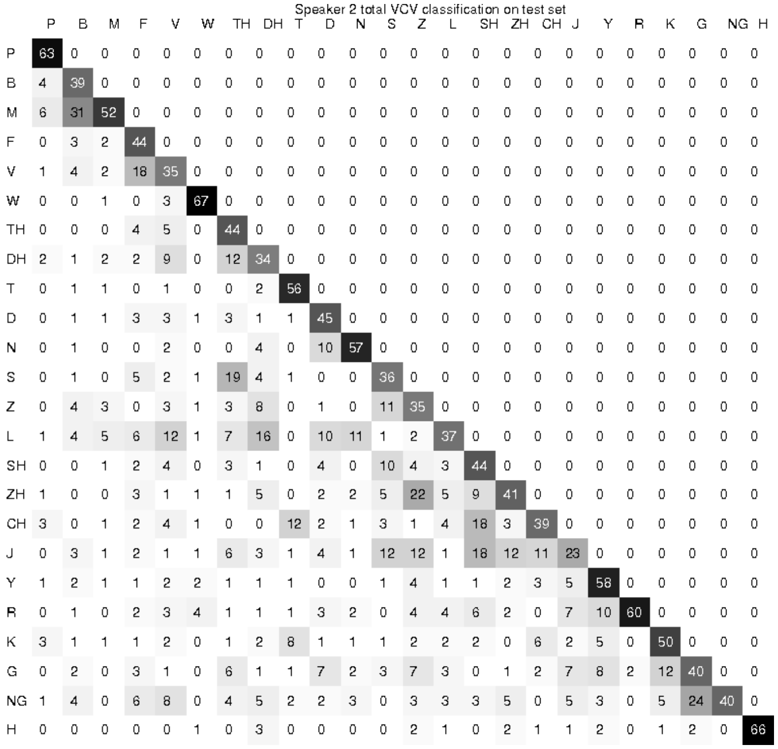
Figure D-2: Simplified confusion matrix of Speaker 1's VCV classifications, with all contexts superimposed.



Speaker 2 vowels classification on test set

	heed	hid	head	had	hod	who'd	heard	hud	hood	hide	how'd	hoed	hoyed	hayed	hawed
heed	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hid	8	12	0	0	0	0	0	0	0	0	0	0	0	0	0
head	2	11	9	0	0	0	0	0	0	0	0	0	0	0	0
had	4	0	1	10	0	0	0	0	0	0	0	0	0	0	0
hod	0	1	1	6	3	0	0	0	0	0	0	0	0	0	0
who'd	0	1	0	0	2	14	0	0	0	0	0	0	0	0	0
heard	0	0	0	2	2	1	14	0	0	0	0	0	0	0	0
hud	0	4	8	0	4	0	0	1	0	0	0	0	0	0	0
hood	0	4	2	0	2	0	0	2	11	0	0	0	0	0	0
hide	0	0	0	0	7	0	0	1	0	9	0	0	0	0	0
how'd	1	0	2	1	2	1	0	1	0	1	14	0	0	0	0
hoed	1	1	0	1	1	4	3	0	0	0	0	12	0	0	0
hoyed	0	0	0	0	0	2	1	1	1	6	0	0	15	0	0
hayed	4	0	0	4	2	0	0	0	0	1	3	1	2	11	0
hawed	0	0	0	6	15	0	0	3	0	0	0	1	0	2	6

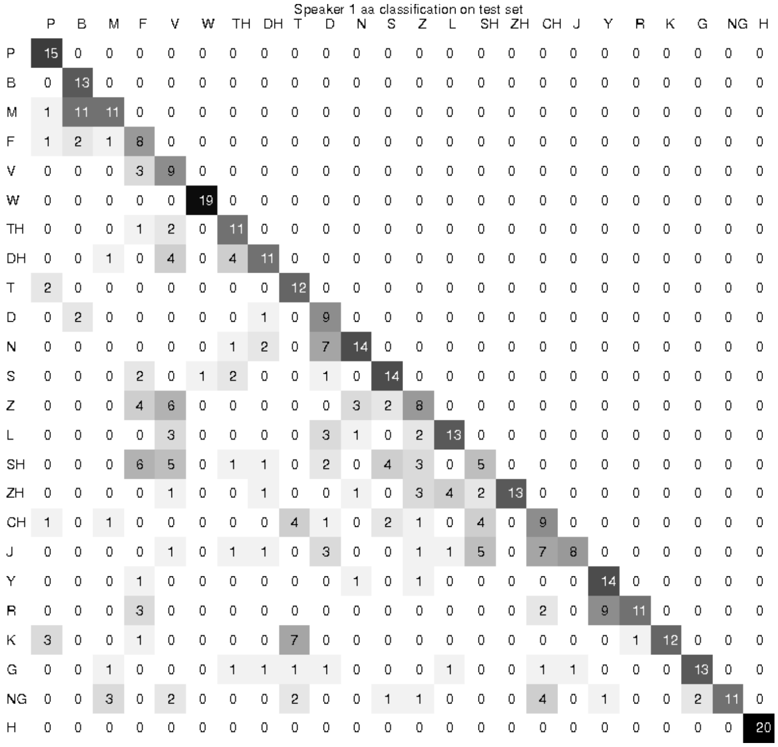
Figure D-3: Simplified confusion matrix of Speaker 2's vowel classification.



Speaker 2 total VCV classification on test set

	P	B	M	F	V	W	TH	DH	T	D	N	S	Z	L	SH	ZH	CH	J	Y	R	K	G	NG	H
P	63	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	4	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	6	31	52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	3	2	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	1	4	2	18	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	1	0	3	67	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TH	0	0	0	4	5	0	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DH	2	1	2	2	9	0	12	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	1	1	0	1	0	0	2	56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	1	1	3	3	1	3	1	1	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	1	0	0	2	0	0	4	0	10	57	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	1	0	5	2	1	19	4	1	0	0	36	0	0	0	0	0	0	0	0	0	0	0	0
Z	0	4	3	0	3	1	3	8	0	1	0	11	35	0	0	0	0	0	0	0	0	0	0	0
L	1	4	5	6	12	1	7	16	0	10	11	1	2	37	0	0	0	0	0	0	0	0	0	0
SH	0	0	1	2	4	0	3	1	0	4	0	10	4	3	44	0	0	0	0	0	0	0	0	0
ZH	1	0	0	3	1	1	1	5	0	2	2	5	22	5	9	41	0	0	0	0	0	0	0	0
CH	3	0	1	2	4	1	0	0	12	2	1	3	1	4	18	3	39	0	0	0	0	0	0	0
J	0	3	1	2	1	1	6	3	1	4	1	12	12	1	18	12	11	23	0	0	0	0	0	0
Y	1	2	1	1	2	2	1	1	1	0	0	1	4	1	1	2	3	5	58	0	0	0	0	0
R	0	1	0	2	3	4	1	1	1	3	2	0	4	4	6	2	0	7	10	60	0	0	0	0
K	3	1	1	1	2	0	1	2	8	1	1	1	2	2	2	0	6	2	5	0	50	0	0	0
G	0	2	0	3	1	0	6	1	1	7	2	3	7	3	0	1	2	7	8	2	12	40	0	0
NG	1	4	0	6	8	0	4	5	2	2	3	0	3	3	3	5	0	5	3	0	5	24	40	0
H	0	0	0	0	0	1	0	3	0	0	0	0	2	1	0	2	1	1	2	0	1	2	0	66

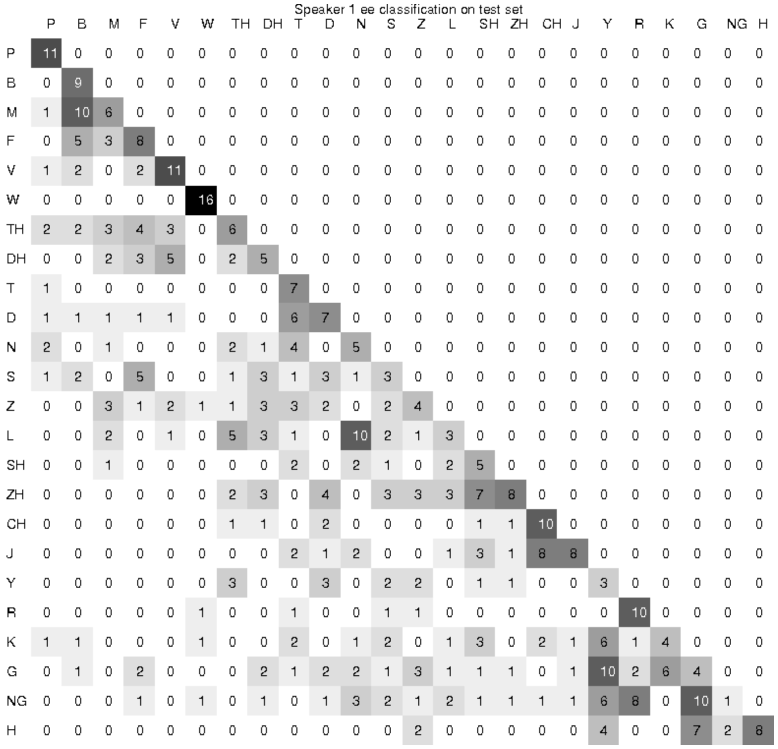
Figure D-4: Simplified confusion matrix of Speaker 2's VCV classifications, with all contexts superimposed.



Speaker 1 aa classification on test set

	P	B	M	F	V	W	TH	DH	T	D	N	S	Z	L	SH	ZH	CH	J	Y	R	K	G	NG	H
P	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	1	11	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	1	2	1	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	0	0	0	3	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TH	0	0	0	1	2	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DH	0	0	1	0	4	0	4	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	2	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	2	0	0	0	0	0	1	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	1	2	0	7	14	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	2	0	1	2	0	0	1	0	14	0	0	0	0	0	0	0	0	0	0	0	0
Z	0	0	0	4	6	0	0	0	0	0	3	2	8	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	3	0	0	0	0	3	1	0	2	13	0	0	0	0	0	0	0	0	0	0
SH	0	0	0	6	5	0	1	1	0	2	0	4	3	0	5	0	0	0	0	0	0	0	0	0
ZH	0	0	0	0	1	0	0	1	0	0	1	0	3	4	2	13	0	0	0	0	0	0	0	0
CH	1	0	1	0	0	0	0	0	4	1	0	2	1	0	4	0	9	0	0	0	0	0	0	0
J	0	0	0	0	1	0	1	1	0	3	0	0	1	1	5	0	7	8	0	0	0	0	0	0
Y	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	14	0	0	0	0	0
R	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	2	0	9	11	0	0	0	0
K	3	0	0	1	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	1	12	0	0	0
G	0	0	1	0	0	0	1	1	1	1	0	0	0	1	0	0	1	1	0	0	0	13	0	0
NG	0	0	3	0	2	0	0	0	2	0	0	1	1	0	0	0	4	0	1	0	0	2	11	0
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20

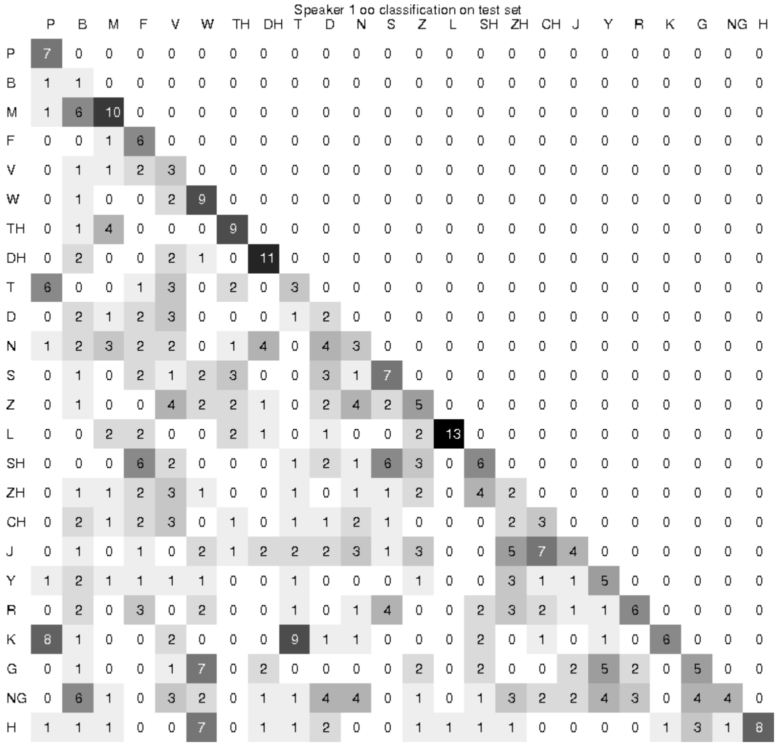
Figure D-5: Simplified confusion matrix of Speaker 1's "aa" context VCV classifications.



Speaker 1 ee classification on test set

	P	B	M	F	V	W	TH	DH	T	D	N	S	Z	L	SH	ZH	CH	J	Y	R	K	G	NG	H
P	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	1	10	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	5	3	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	1	2	0	2	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TH	2	2	3	4	3	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DH	0	0	2	3	5	0	2	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	1	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	1	1	1	1	1	0	0	0	6	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	2	0	1	0	0	0	2	1	4	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
S	1	2	0	5	0	0	1	3	1	3	1	3	0	0	0	0	0	0	0	0	0	0	0	0
Z	0	0	3	1	2	1	1	3	3	2	0	2	4	0	0	0	0	0	0	0	0	0	0	0
L	0	0	2	0	1	0	5	3	1	0	10	2	1	3	0	0	0	0	0	0	0	0	0	0
SH	0	0	1	0	0	0	0	0	2	0	2	1	0	2	5	0	0	0	0	0	0	0	0	0
ZH	0	0	0	0	0	0	2	3	0	4	0	3	3	3	7	8	0	0	0	0	0	0	0	0
CH	0	0	0	0	0	0	1	1	0	2	0	0	0	0	1	1	10	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	2	1	2	0	0	1	3	1	8	8	0	0	0	0	0	0
Y	0	0	0	0	0	0	3	0	0	3	0	2	2	0	1	1	0	0	3	0	0	0	0	0
R	0	0	0	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	10	0	0	0	0
K	1	1	0	0	0	1	0	0	2	0	1	2	0	1	3	0	2	1	6	1	4	0	0	0
G	0	1	0	2	0	0	0	2	1	2	2	1	3	1	1	1	0	1	10	2	6	4	0	0
NG	0	0	0	1	0	1	0	1	0	1	3	2	1	2	1	1	1	1	6	8	0	10	1	0
H	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	4	0	0	7	2	8

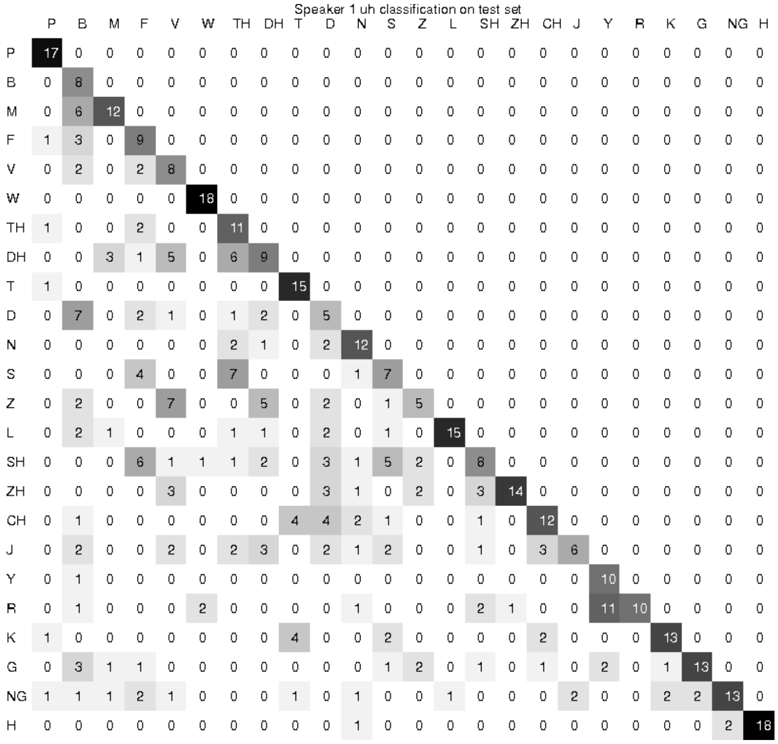
Figure D-6: Simplified confusion matrix of Speaker 1's "ee" context VCV classifications.



Speaker 1 oo classification on test set

	P	B	M	F	V	W	TH	DH	T	D	N	S	Z	L	SH	ZH	CH	J	Y	R	K	G	NG	H
P	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	1	6	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	1	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	0	1	1	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	1	0	0	2	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TH	0	1	4	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DH	0	2	0	0	2	1	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	6	0	0	1	3	0	2	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	2	1	2	3	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	1	2	3	2	2	0	1	4	0	4	3	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	1	0	2	1	2	3	0	0	3	1	7	0	0	0	0	0	0	0	0	0	0	0	0
Z	0	1	0	0	4	2	2	1	0	2	4	2	5	0	0	0	0	0	0	0	0	0	0	0
L	0	0	2	2	0	0	2	1	0	1	0	0	2	13	0	0	0	0	0	0	0	0	0	0
SH	0	0	0	6	2	0	0	0	1	2	1	6	3	0	6	0	0	0	0	0	0	0	0	0
ZH	0	1	1	2	3	1	0	0	1	0	1	1	2	0	4	2	0	0	0	0	0	0	0	0
CH	0	2	1	2	3	0	1	0	1	1	2	1	0	0	0	2	3	0	0	0	0	0	0	0
J	0	1	0	1	0	2	1	2	2	2	3	1	3	0	0	5	7	4	0	0	0	0	0	0
Y	1	2	1	1	1	1	0	0	1	0	0	0	1	0	0	3	1	1	5	0	0	0	0	0
R	0	2	0	3	0	2	0	0	1	0	1	4	0	0	2	3	2	1	1	6	0	0	0	0
K	8	1	0	0	2	0	0	0	9	1	1	0	0	0	2	0	1	0	1	0	6	0	0	0
G	0	1	0	0	1	7	0	2	0	0	0	0	2	0	2	0	0	2	5	2	0	5	0	0
NG	0	6	1	0	3	2	0	1	1	4	4	0	1	0	1	3	2	2	4	3	0	4	4	0
H	1	1	1	0	0	7	0	1	1	2	0	0	1	1	1	1	0	0	0	0	1	3	1	8

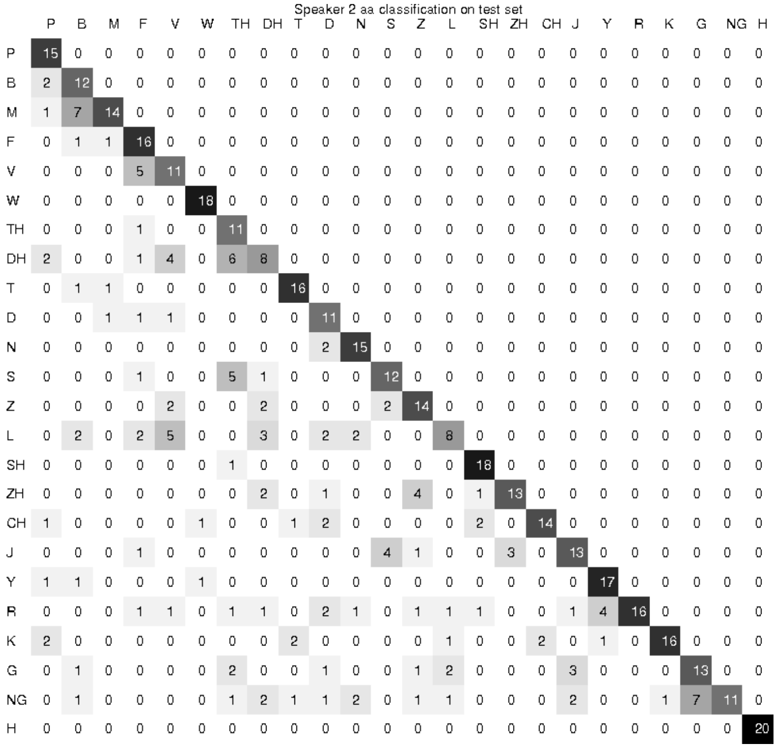
Figure D-7: Simplified confusion matrix of Speaker 1's "oo" context VCV classifications.



Speaker 1 uh classification on test set

	P	B	M	F	V	W	TH	DH	T	D	N	S	Z	L	SH	ZH	CH	J	Y	R	K	G	NG	H
P	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	6	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	1	3	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	0	2	0	2	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TH	1	0	0	2	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DH	0	0	3	1	5	0	6	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	1	0	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	7	0	2	1	0	1	2	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	2	1	0	2	12	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	4	0	0	7	0	0	0	1	7	0	0	0	0	0	0	0	0	0	0	0	0
Z	0	2	0	0	7	0	0	5	0	2	0	1	5	0	0	0	0	0	0	0	0	0	0	0
L	0	2	1	0	0	0	1	1	0	2	0	1	0	15	0	0	0	0	0	0	0	0	0	0
SH	0	0	0	6	1	1	1	2	0	3	1	5	2	0	8	0	0	0	0	0	0	0	0	0
ZH	0	0	0	0	3	0	0	0	0	3	1	0	2	0	3	14	0	0	0	0	0	0	0	0
CH	0	1	0	0	0	0	0	0	4	4	2	1	0	0	1	0	12	0	0	0	0	0	0	0
J	0	2	0	0	2	0	2	3	0	2	1	2	0	0	1	0	3	6	0	0	0	0	0	0
Y	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0
R	0	1	0	0	0	2	0	0	0	0	1	0	0	0	2	1	0	0	11	10	0	0	0	0
K	1	0	0	0	0	0	0	0	4	0	0	2	0	0	0	0	2	0	0	0	13	0	0	0
G	0	3	1	1	0	0	0	0	0	0	0	1	2	0	1	0	1	0	2	0	1	13	0	0
NG	1	1	1	2	1	0	0	0	1	0	1	0	0	1	0	0	0	2	0	0	2	2	13	0
H	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	18

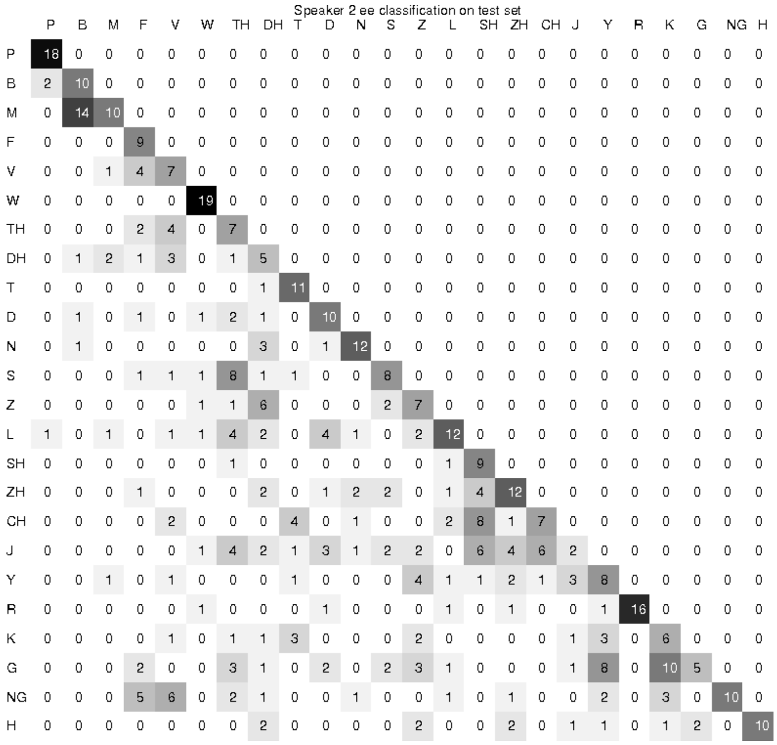
Figure D-8: Simplified confusion matrix of Speaker 1's "uh" context VCV classifications.



Speaker 2 aa classification on test set

	P	B	M	F	V	W	TH	DH	T	D	N	S	Z	L	SH	ZH	CH	J	Y	R	K	G	NG	H
P	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	2	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	1	7	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	1	1	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	0	0	0	5	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TH	0	0	0	1	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DH	2	0	0	1	4	0	6	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	1	1	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	1	1	1	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	2	15	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	1	0	0	5	1	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0
Z	0	0	0	0	2	0	0	2	0	0	0	2	14	0	0	0	0	0	0	0	0	0	0	0
L	0	2	0	2	5	0	0	3	0	2	2	0	0	8	0	0	0	0	0	0	0	0	0	0
SH	0	0	0	0	0	0	1	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0
ZH	0	0	0	0	0	0	0	2	0	1	0	0	4	0	1	13	0	0	0	0	0	0	0	0
CH	1	0	0	0	0	1	0	0	1	2	0	0	0	0	2	0	14	0	0	0	0	0	0	0
J	0	0	0	1	0	0	0	0	0	0	0	4	1	0	0	3	0	13	0	0	0	0	0	0
Y	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0
R	0	0	0	1	1	0	1	1	0	2	1	0	1	1	1	0	0	1	4	16	0	0	0	0
K	2	0	0	0	0	0	0	0	2	0	0	0	0	1	0	0	2	0	1	0	16	0	0	0
G	0	1	0	0	0	0	2	0	0	1	0	0	1	2	0	0	0	3	0	0	0	13	0	0
NG	0	1	0	0	0	0	1	2	1	1	2	0	1	1	0	0	0	2	0	0	1	7	11	0
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20

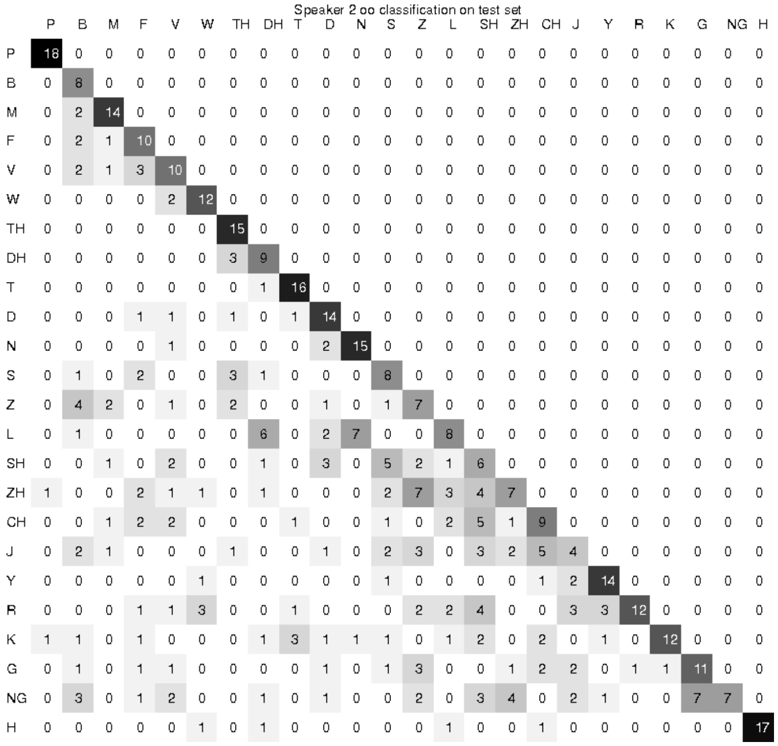
Figure D-9: Simplified confusion matrix of Speaker 2's "aa" context VCV classifications.



Speaker 2 ee classification on test set

	P	B	M	F	V	W	TH	DH	T	D	N	S	Z	L	SH	ZH	CH	J	Y	R	K	G	NG	H
P	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	2	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	14	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	0	0	1	4	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TH	0	0	0	2	4	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DH	0	1	2	1	3	0	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	1	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	1	0	1	0	1	2	1	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	1	0	0	0	0	0	3	0	1	12	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	1	1	1	8	1	1	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0
Z	0	0	0	0	0	1	1	6	0	0	0	2	7	0	0	0	0	0	0	0	0	0	0	0
L	1	0	1	0	1	1	4	2	0	4	1	0	2	12	0	0	0	0	0	0	0	0	0	0
SH	0	0	0	0	0	0	1	0	0	0	0	0	0	1	9	0	0	0	0	0	0	0	0	0
ZH	0	0	0	1	0	0	0	2	0	1	2	2	0	1	4	12	0	0	0	0	0	0	0	0
CH	0	0	0	0	2	0	0	0	4	0	1	0	0	2	8	1	7	0	0	0	0	0	0	0
J	0	0	0	0	0	1	4	2	1	3	1	2	2	0	6	4	6	2	0	0	0	0	0	0
Y	0	0	1	0	1	0	0	0	1	0	0	0	4	1	1	2	1	3	8	0	0	0	0	0
R	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	1	0	0	1	16	0	0	0	0
K	0	0	0	0	1	0	1	1	3	0	0	0	2	0	0	0	0	1	3	0	6	0	0	0
G	0	0	0	2	0	0	3	1	0	2	0	2	3	1	0	0	0	1	8	0	10	5	0	0
NG	0	0	0	5	6	0	2	1	0	0	1	0	0	1	0	1	0	0	2	0	3	0	10	0
H	0	0	0	0	0	0	0	2	0	0	0	0	2	0	0	2	0	1	1	0	1	2	0	10

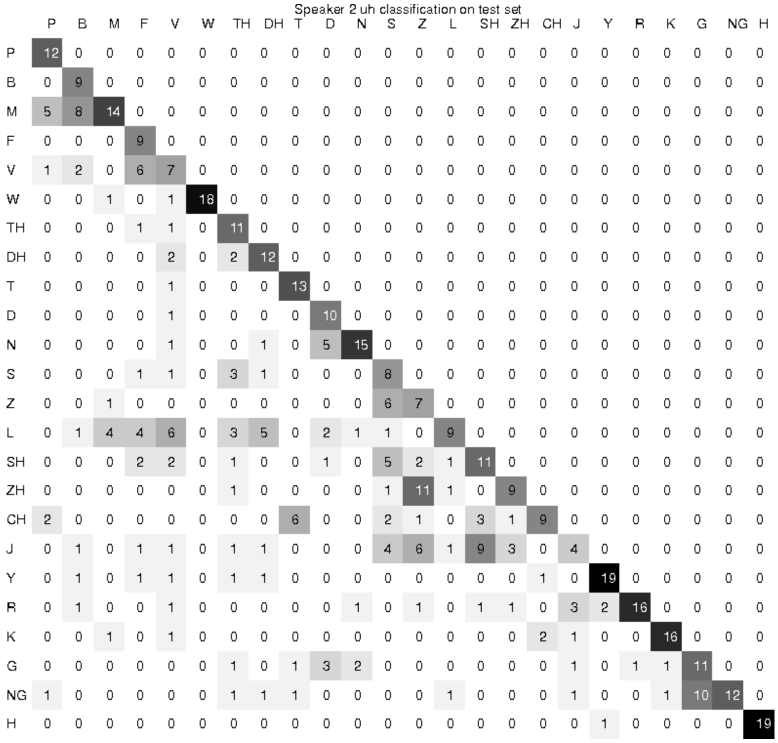
Figure D-10: Simplified confusion matrix of Speaker 2's "ee" context VCV classifications.



Speaker 2 oo classification on test set

	P	B	M	F	V	W	TH	DH	T	D	N	S	Z	L	SH	ZH	CH	J	Y	R	K	G	NG	H
P	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	2	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	2	1	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	0	2	1	3	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	2	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TH	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DH	0	0	0	0	0	0	3	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	1	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	1	1	0	1	0	1	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	1	0	0	0	0	2	15	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	1	0	2	0	0	3	1	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0
Z	0	4	2	0	1	0	2	0	0	1	0	1	7	0	0	0	0	0	0	0	0	0	0	0
L	0	1	0	0	0	0	0	6	0	2	7	0	0	8	0	0	0	0	0	0	0	0	0	0
SH	0	0	1	0	2	0	0	1	0	3	0	5	2	1	6	0	0	0	0	0	0	0	0	0
ZH	1	0	0	2	1	1	0	1	0	0	0	2	7	3	4	7	0	0	0	0	0	0	0	0
CH	0	0	1	2	2	0	0	0	1	0	0	1	0	2	5	1	9	0	0	0	0	0	0	0
J	0	2	1	0	0	0	1	0	0	1	0	2	3	0	3	2	5	4	0	0	0	0	0	0
Y	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	2	14	0	0	0	0	0
R	0	0	0	1	1	3	0	0	1	0	0	0	2	2	4	0	0	3	3	12	0	0	0	0
K	1	1	0	1	0	0	0	1	3	1	1	1	0	1	2	0	2	0	1	0	12	0	0	0
G	0	1	0	1	1	0	0	0	0	1	0	1	3	0	0	1	2	2	0	1	1	11	0	0
NG	0	3	0	1	2	0	0	1	0	1	0	0	2	0	3	4	0	2	1	0	0	7	7	0
H	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	17

Figure D-11: Simplified confusion matrix of Speaker 2's "oo" context VCV classifications.



Speaker 2 uh classification on test set

	P	B	M	F	V	W	TH	DH	T	D	N	S	Z	L	SH	ZH	CH	J	Y	R	K	G	NG	H
P	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	5	8	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	1	2	0	6	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	1	0	1	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TH	0	0	0	1	1	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DH	0	0	0	0	2	0	2	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	1	0	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	1	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	1	0	0	1	0	5	15	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	1	1	0	3	1	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0
Z	0	0	1	0	0	0	0	0	0	0	0	6	7	0	0	0	0	0	0	0	0	0	0	0
L	0	1	4	4	6	0	3	5	0	2	1	1	0	9	0	0	0	0	0	0	0	0	0	0
SH	0	0	0	2	2	0	1	0	0	1	0	5	2	1	11	0	0	0	0	0	0	0	0	0
ZH	0	0	0	0	0	0	1	0	0	0	0	1	11	1	0	9	0	0	0	0	0	0	0	0
CH	2	0	0	0	0	0	0	0	6	0	0	2	1	0	3	1	9	0	0	0	0	0	0	0
J	0	1	0	1	1	0	1	1	0	0	0	4	6	1	9	3	0	4	0	0	0	0	0	0
Y	0	1	0	1	1	0	1	1	0	0	0	0	0	0	0	0	1	0	19	0	0	0	0	0
R	0	1	0	0	1	0	0	0	0	0	1	0	1	0	1	1	0	3	2	16	0	0	0	0
K	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	16	0	0	0
G	0	0	0	0	0	0	1	0	1	3	2	0	0	0	0	0	0	1	0	1	1	11	0	0
NG	1	0	0	0	0	0	1	1	1	0	0	0	0	1	0	0	0	1	0	0	1	10	12	0
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	19

Figure D-12: Simplified confusion matrix of Speaker 2's "uh" context VCV classifications.

Appendix E

Tables of Highly Confusable Consonant Pairs

Speaker 1 (F) Top 10 Confusions - Consonants								
“aa”			“ee”		“oo”		“uh”	
Rank	Pair	Miscl. %	Pair	Miscl. %	Pair	Miscl. %	Pair	Miscl. %
1	M-B	27.50	M-B	25.00	K-T	22.50	R-Y	27.50
2	R-Y	22.50	L-N	25.00	K-P	20.00	D-B	17.50
3	K-T	17.50	G-Y	25.00	G-W	17.50	Z-V	17.50
4	N-D	17.50	NG-G	25.00	H-W	17.50	S-TH	17.50
5	J-CH	17.50	J-CH	20.00	J-CH	17.50	M-B	15.00
6	SH-F	15.00	NG-R	20.00	T-P	15.00	SH-F	15.00
7	Z-V	15.00	ZH-SH	17.50	M-B	15.00	DH-TH	15.00
8	SH-V	12.50	H-G	17.50	NG-B	15.00	DH-V	12.50
9	J-SH	12.50	D-T	15.00	SH-F	15.00	Z-DH	12.50
10	Z-F	10.00	K-Y	15.00	SH-S	15.00	SH-S	12.50

Table E.1: Speaker 1 Top 10 Misclassified Pairs of VCVs in each context.

Speaker 2 (M) Top 10 Confusions - Consonants								
“aa”			“ee”		“oo”		“uh”	
Rank	Pair	Miscl. %	Pair	Miscl. %	Pair	Miscl. %	Pair	Miscl. %
1	M-B	17.50	M-B	35.00	L-N	17.50	ZH-Z	27.50
2	NG-G	17.50	G-K	25.00	ZH-Z	17.50	NG-G	25.00
3	DH-TH	15.00	S-TH	20.00	NG-G	17.50	J-SH	22.50
4	V-F	12.50	CH-SH	20.00	L-DH	15.00	M-B	20.00
5	L-V	12.50	G-Y	20.00	SH-S	12.50	V-F	15.00
6	S-TH	12.50	NG-V	15.00	CH-SH	12.50	L-V	15.00
7	DH-V	10.00	Z-DH	15.00	J-CH	12.50	CH-T	15.00
8	J-S	10.00	J-SH	15.00	Z-B	10.00	Z-S	15.00
9	ZH-Z	10.00	J-CH	15.00	ZH-SH	10.00	J-Z	15.00
10	R-Y	10.00	NG-F	12.50	R-SH	10.00	M-P	12.50

Table E.2: Speaker 2 Top 10 Misclassified Pairs of VCVs in each context.

Appendix F

Spectrograms of Phone Classification Data

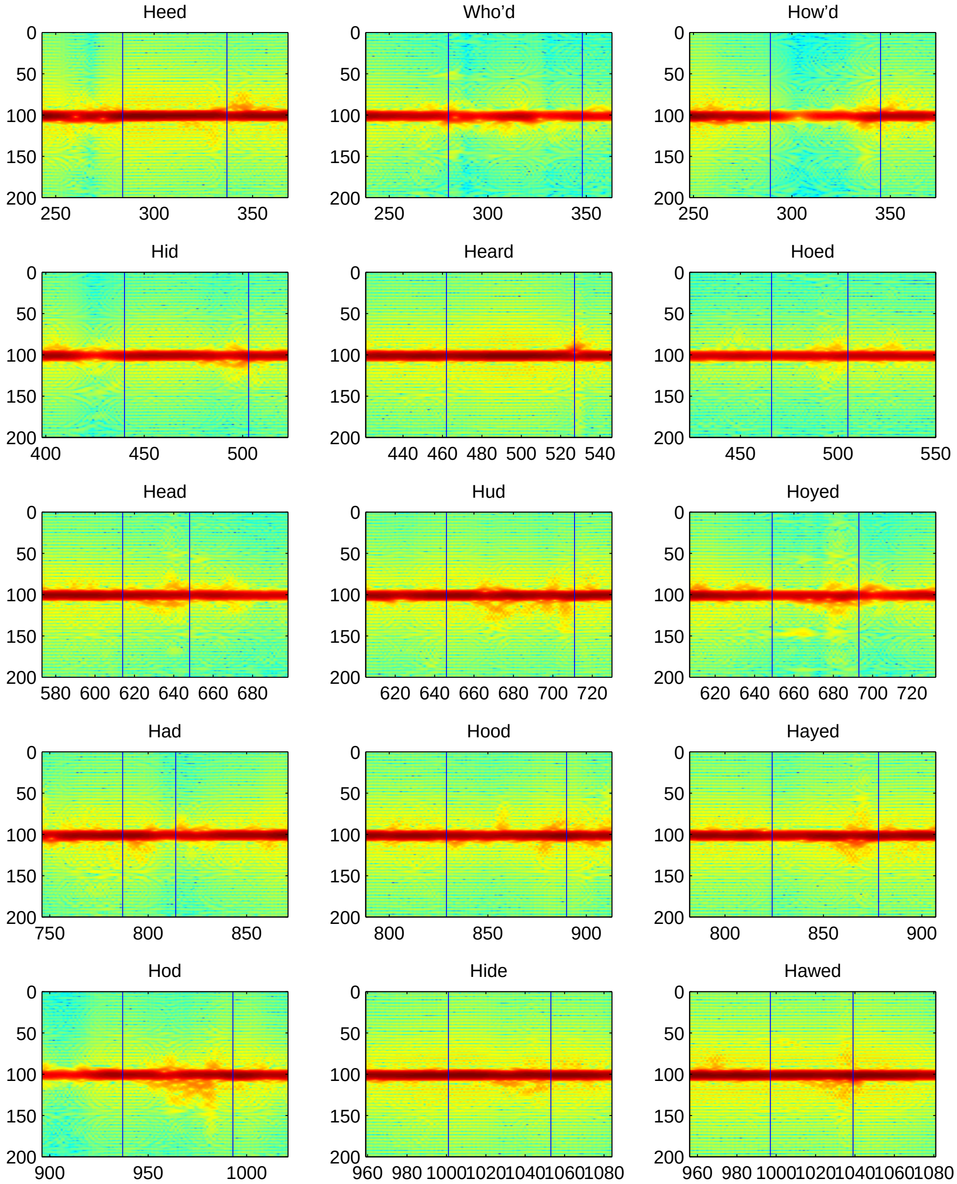


Figure F-1: Spectrograms of Speaker 1 vowels. Time (in 6ms frames) is represented by the x-axis, while frequency (in 4 Hz frames) is represented by the y-axis.

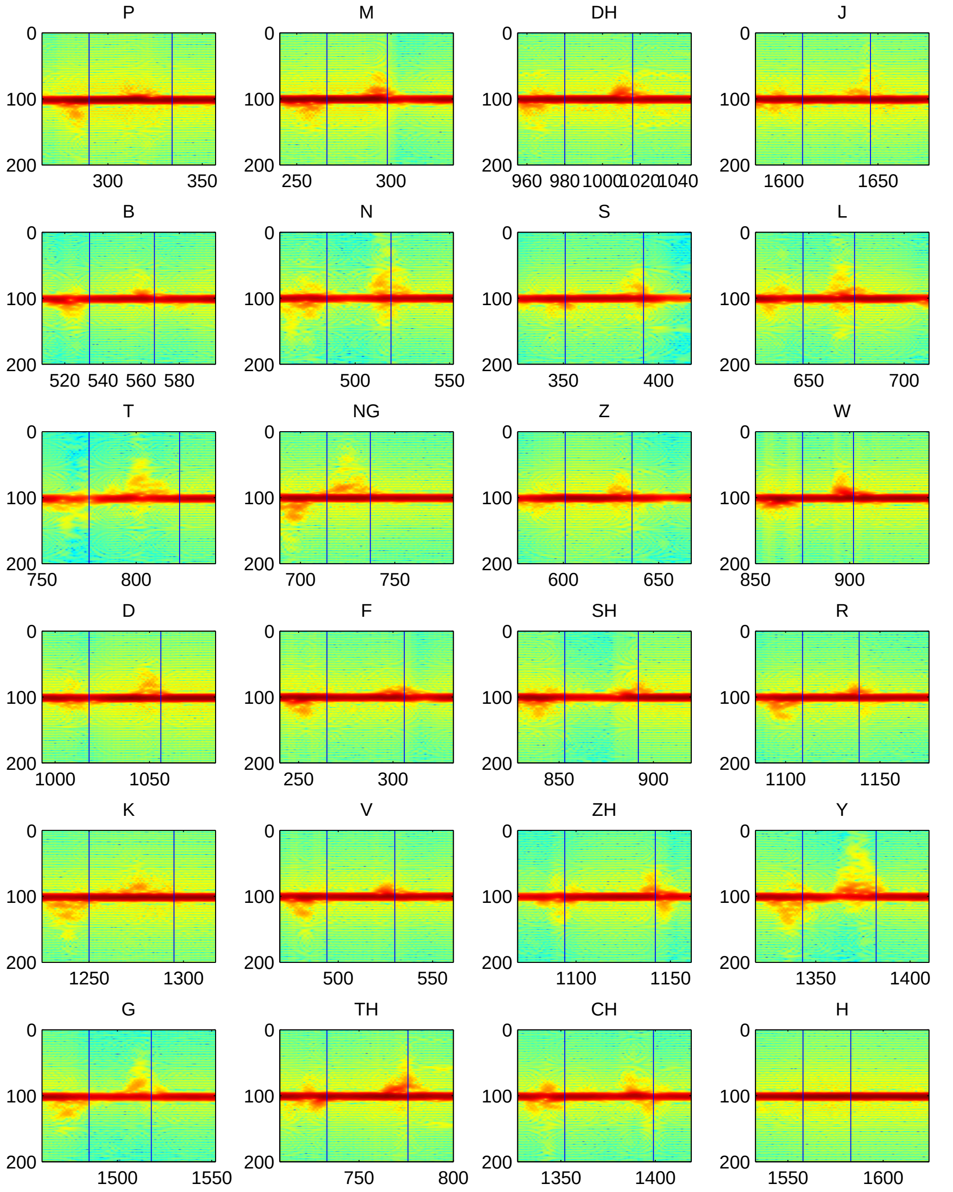


Figure F-2: Spectrograms of Speaker 1 “aa” context VCVs. Time (in 6ms frames) is represented by the x-axis, while frequency (in 4 Hz frames) is represented by the y-axis.

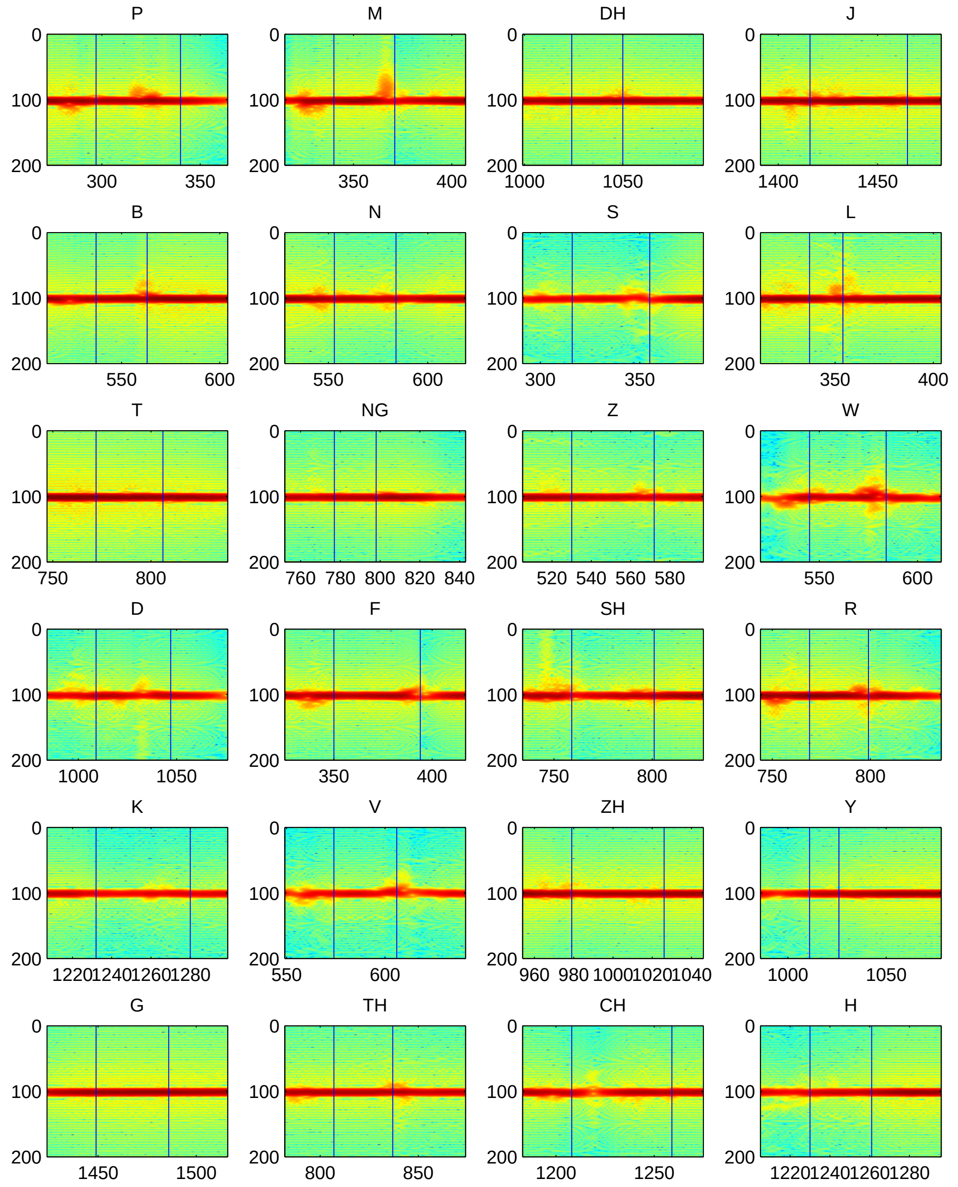


Figure F-3: Spectrograms of Speaker 1 “ee” context VCVs. Time (in 6ms frames) is represented by the x-axis, while frequency (in 4 Hz frames) is represented by the y-axis.

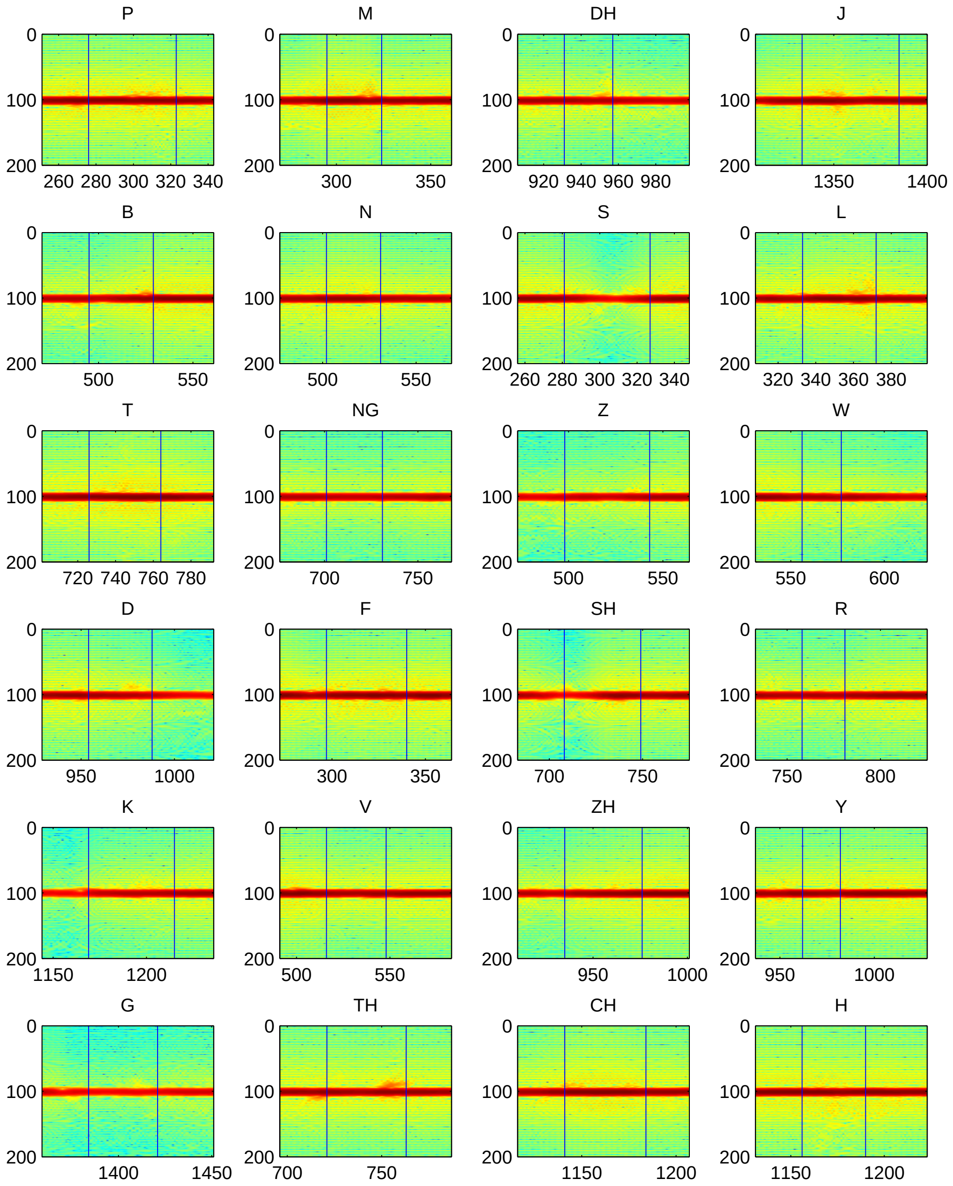


Figure F-4: Spectrograms of Speaker 1 "oo" context VCVs. Time (in 6ms frames) is represented by the x-axis, while frequency (in 4 Hz frames) is represented by the y-axis.

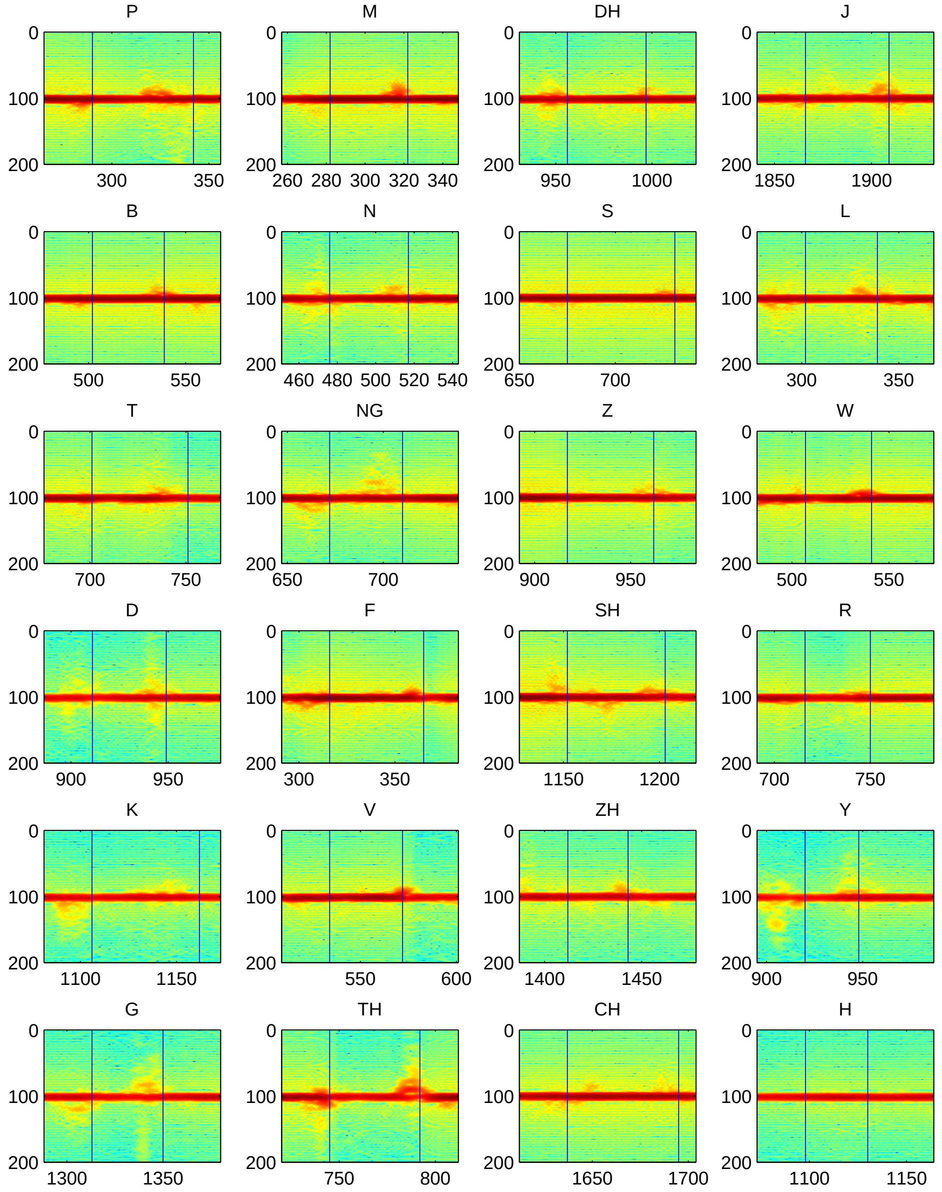


Figure F-5: Spectrograms of Speaker 1 “uh” context VCVs. Time (in 6ms frames) is represented by the x-axis, while frequency (in 4 Hz frames) is represented by the y-axis.

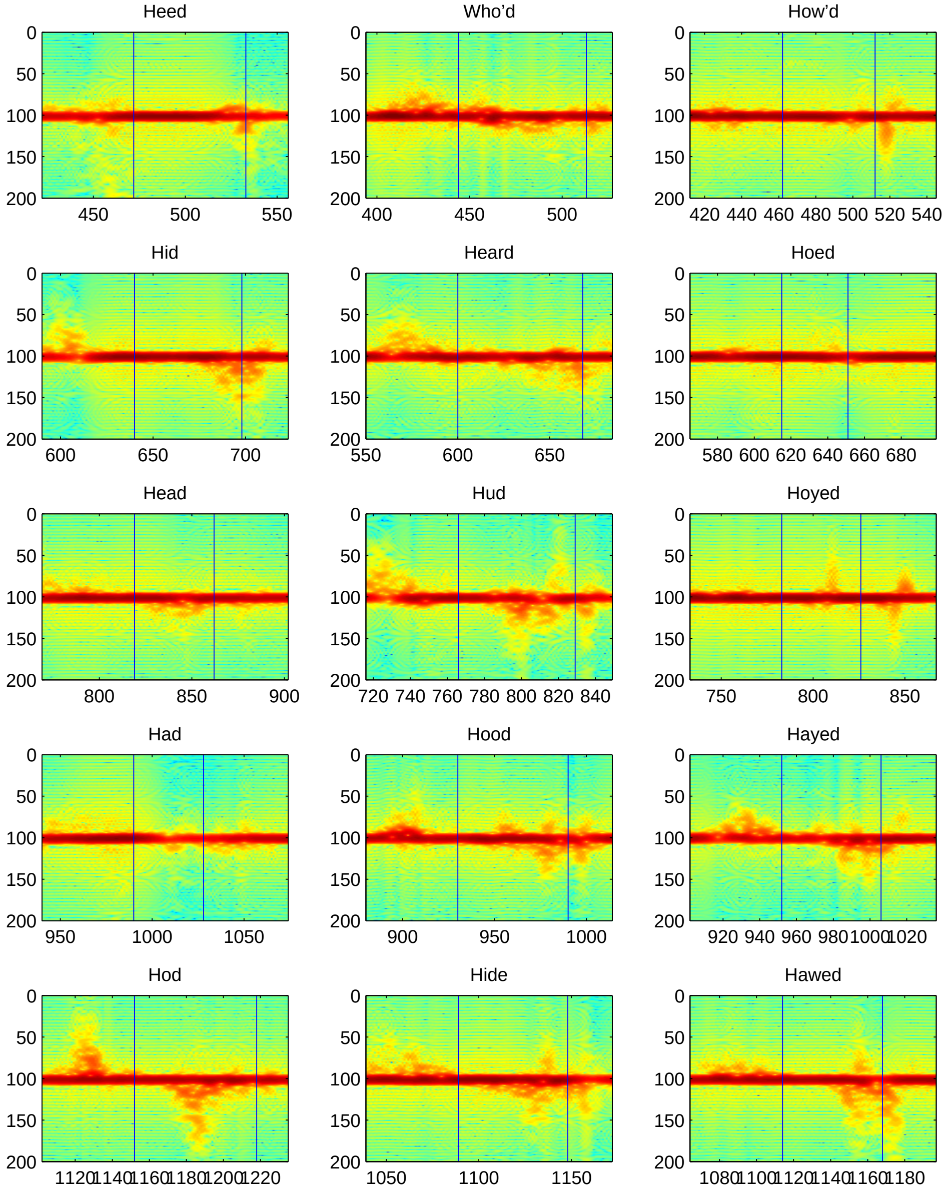


Figure F-6: Spectrograms of Speaker 2 vowels. Time (in 6ms frames) is represented by the x-axis, while frequency (in 4 Hz frames) is represented by the y-axis.

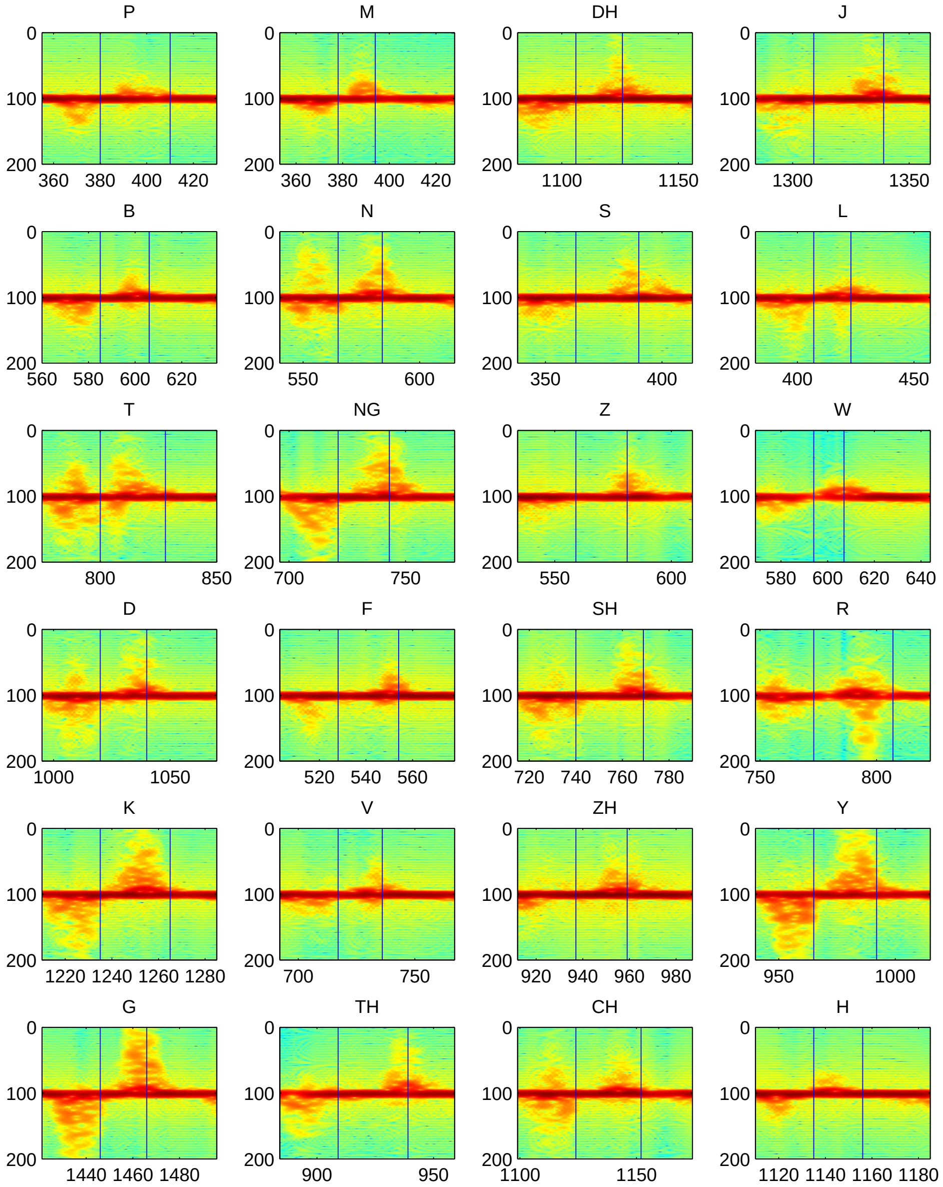


Figure F-7: Spectrograms of Speaker 2 "aa" context VCVs. Time (in 6ms frames) is represented by the x-axis, while frequency (in 4 Hz frames) is represented by the y-axis.

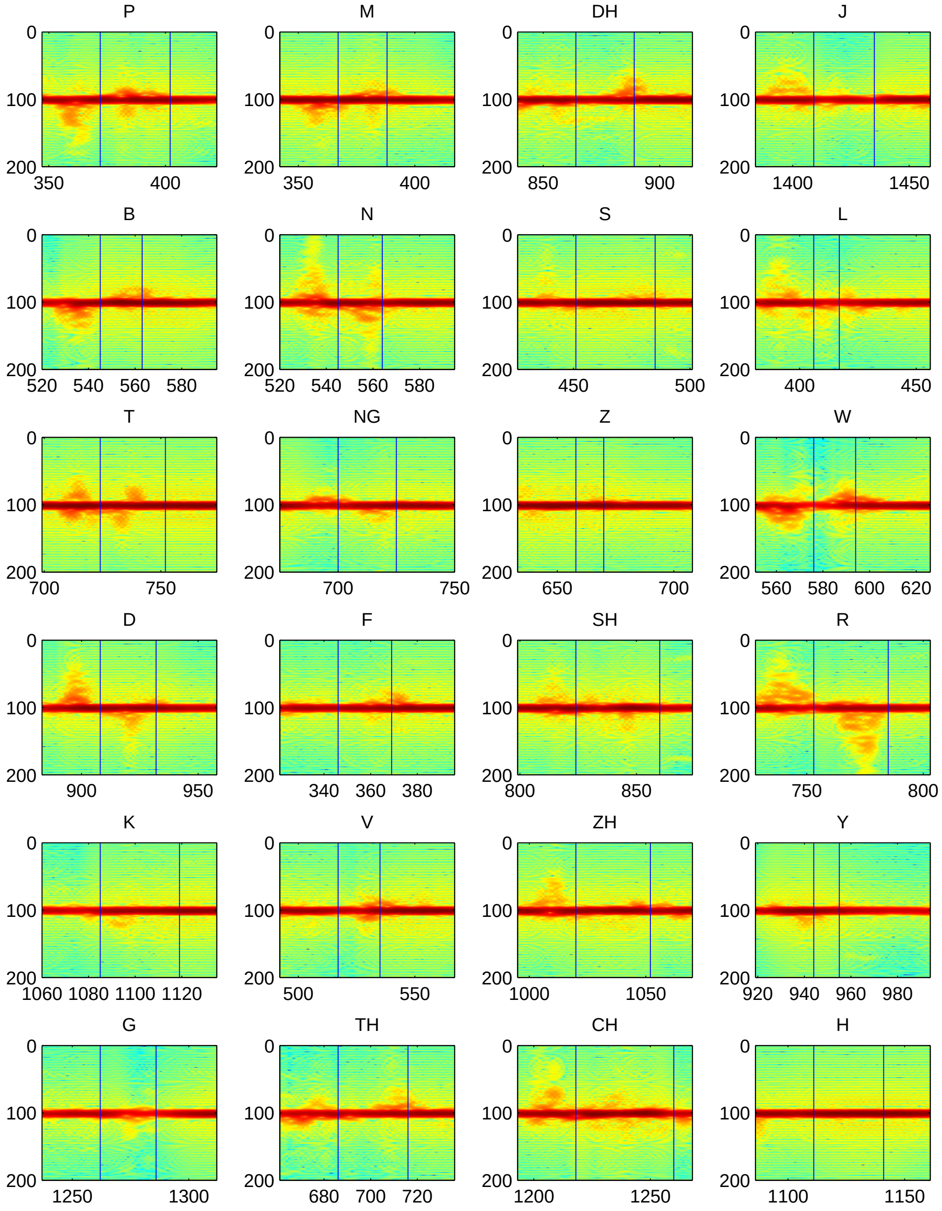


Figure F-8: Spectrograms of Speaker 2 "ee" context VCVs. Time (in 6ms frames) is represented by the x-axis, while frequency (in 4 Hz frames) is represented by the y-axis.

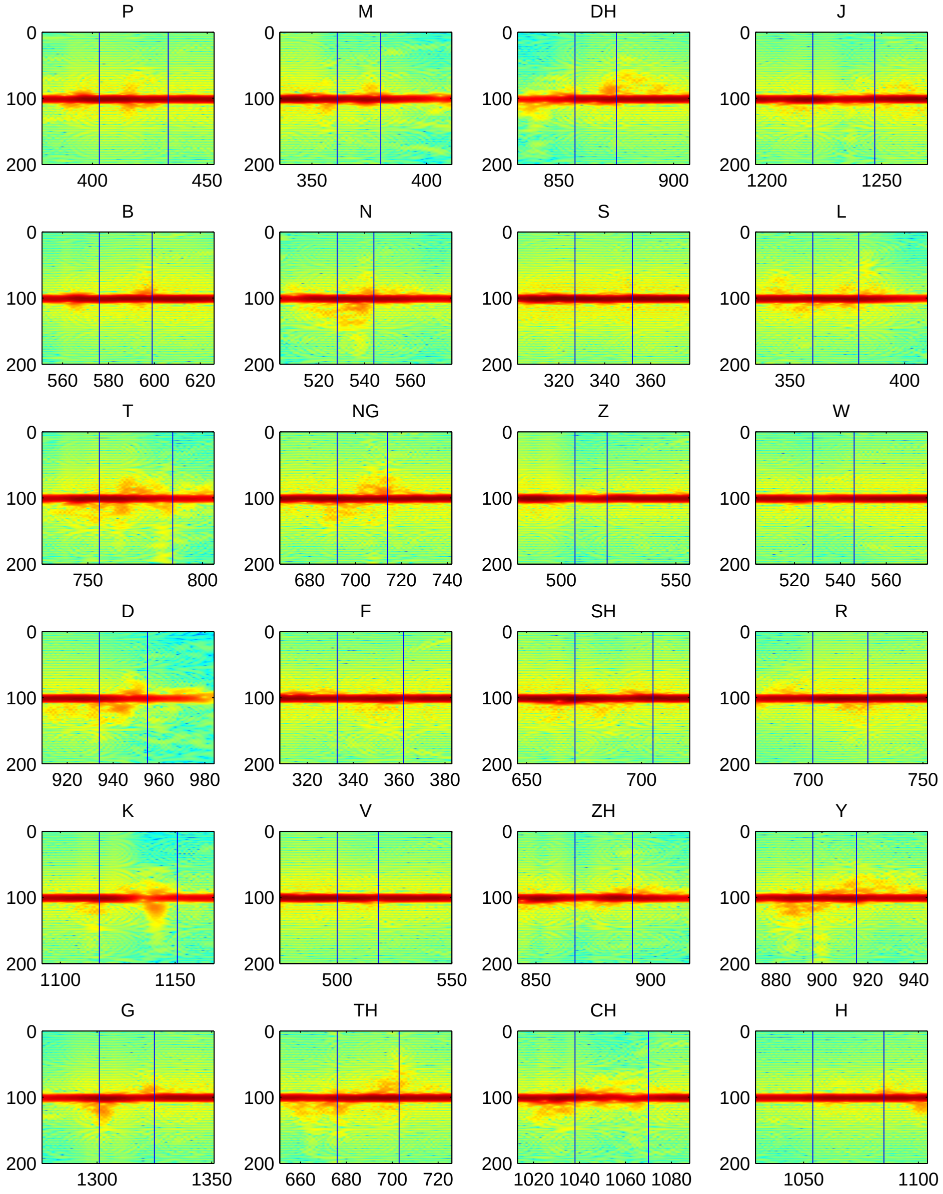


Figure F-9: Spectrograms of Speaker 2 “oo” context VCVs. Time (in 6ms frames) is represented by the x-axis, while frequency (in 4 Hz frames) is represented by the y-axis.

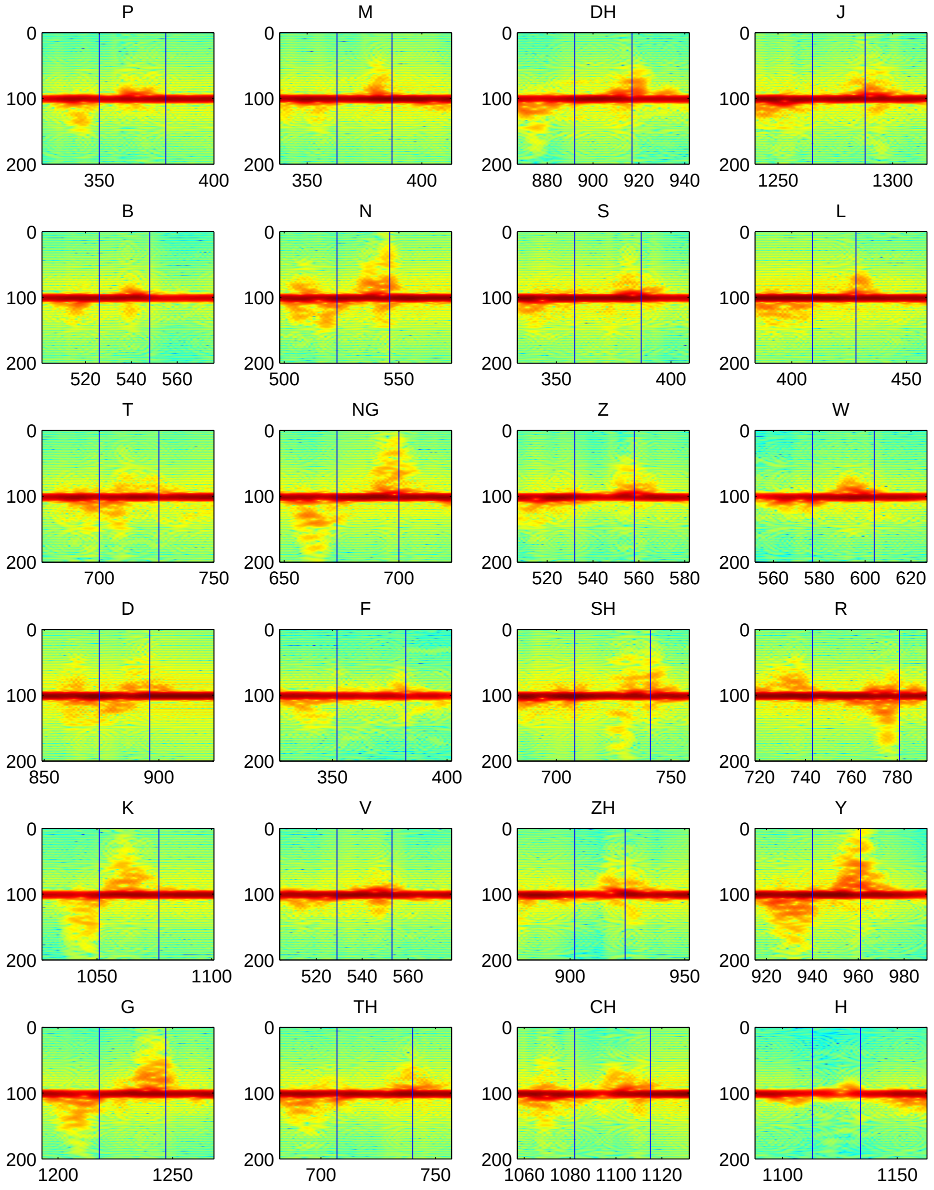


Figure F-10: Spectrograms of Speaker 2 “uh” context VCVs. Time (in 6ms frames) is represented by the x-axis, while frequency (in 4 Hz frames) is represented by the y-axis.

Bibliography

[1] K. Aizawa, Y. Nakamura, and S. Satoh. *Advances in Multimedia Information Processing: PCM 2004*, chapter 3.1. Springer, 2004.

[2] L. Bernstein and C. Benoit. For speech perception by humans or machines, three senses are better than one. *Proc. ICSLP*, 1996.

[3] M. Breeuwer and R. Plomp. Speechreading supplemented with auditorily presented speech parameters. *J. Acoust. Soc. Am.*, 79:481–499, 1986.

[4] S. Chu and T. Huang. Audio-visual speech modeling using coupled hidden Markov models. *Proc. ICASSP*, 2002.

[5] M. J. F. Gales. *Model-based techniques for noise robust speech recognition.* PhD thesis, University of Cambridge, 1995.

[6] J. Glass. A probabilistic framework for segment-based speech recognition. *Computer, Speech, and Language*, 17(2-3):137–152, 2003.

[7] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt. Combining standard and throat microphones for robust speech recognition. *IEEE Signal Processing Letters*, 10(3):72–74, 2003.

[8] T.J. Hazen. Visual model structures and synchrony constraints for audio-visual speech recognition. *IEEE Trans. Audio, Speech, and Language Processing*, 14(3):1082–1089, 2006.

[9] P. Holmberg. Robust ultrasonic range finder - an FFT analysis. *Measurement Science and Technology*, 1992.

[10] D.L. Jennings and D.W. Ruck. Enhancing automatic speech recognition with an ultrasonic lipmotion detector. *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 1995.

[11] K. Kalgaonkar and B. Raj. An acoustic doppler-based front-end for hands free spoken user interfaces. *IEEE/ACL Workshop on Spoken Language Technology*, 2006.

[12] K. Kalgaonkar and B. Raj. Ultrasonic doppler sensor for speaker recognition. *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2008.

[13] C. Kwan, X. Li, D. Law, Y. Deng, Z. Ren, B. Raj, R. Singh, and R. Stern. Voice driven applications in non-stationary and chaotic environment. *Trans. International Journal of Signal Processing*, 3(4), 2006.

[14] H. McGurk and J. MacDonald. Hearing lips and seeing voices: A new illusion. *Nature*, 264:746–748, 1976.

[15] T.R. Nelson. Three-dimensional ultrasound imaging. *Ultrasound in Medicine and Biology*, 24(9):1243–70, 1998.

[16] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. Audio-visual speech recognition. *Center for Language and Speech Processing Final Workshop 2000 Report*, 2000.

[17] L. Neumeyer and M. Weintraub. Probabilistic optimum filtering for robust speech recognition. *Proc. ICASSP*, 1994.

[18] L. C. Ng, J. F. Holzrichter, and T. J. Gable. Denoising of human speech using combined acoustic and EM sensor signal processing. *Proc. ICASSP*, 2000.

[19] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proc. IEEE*, 91(9):1306–1326, 2003.

[20] A.Q. Summerfield. Some preliminaries to a comprehensive account of audio-visual speech perception. In *Hearing by Eye: The Psychology of Lip-Reading*, pages 3–51. Lawrence Erlbaum Associates, London, 1987.

[21] S. Tamura, K. Iwano, and S. Furui. A stream-weight optimization method for audio-visual speech recognition using multi-stream HMMs. *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2004.

[22] P. Teissier, J. Robert-Ribes, and J. L. Schwartz. Comparing models for audiovisual fusion in a noisy-vowel recognition task. *IEEE Trans. Speech Audio Processing*, 7:629–642, 1999.

[23] A. Varga and H. Steeneken. Assessment for automatic speech recognition ii: Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251, 1993.

[24] Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo, X. Huang, and Y. Zheng. Multi-sensory microphones for robust speech detection, enhancement and recognition. *Proc. IEEE*, 3:781–4, 2004.

[25] B. Zhu, T. J. Hazen, and J. Glass. Multimodal speech recognition with ultrasonic sensors. *Proc. Interspeech*, 2007.