VATA: AN IMPROVED PERSONAL COMPUTER-BASED VOWEL ARTICULATION TRAINING AID

by

Andrew Matthew Zimmer
B.A. May 1992, University of Virginia
B.S. May 1997, Old Dominion University

A Thesis Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirement for the Degree of

MASTER OF SCIENCE

ELECTRICAL ENGINEERING

OLD DOMINION UNIVERSITY
May 2002

Approved by:

____
Stephen A. Zahorian (Director)

____
Vijayan K. Asari (Member)

____
Ravindra P. Joshi (Member)

ABSTRACT

VATA: AN IMPROVED PERSONAL COMPUTER-BASED VOWEL ARTICULATION TRAINING AID

Andrew Matthew Zimmer
Old Dominion University, 2002
Director: Dr. Stephen A. Zahorian

This thesis discusses work accomplished on the further development of a vowel articulation training aid for persons with hearing impairments. The system produces visual feedback about the quality of articulation for ten American English monophthong vowel phonemes. A large database of vowel recordings has been collected and used to improve and test the recognition rates of artificial neural networks used by the training aid. A maximum-likelihood classifier has been introduced into the system to improve performance by reducing the likelihood of incorrect feedback produced by the displays. A series of experiments has been performed to examine effects of tuning parameters, and of varying the amount and type of speech data used to train the system. Much of the program code for the speech display system has been revised to use a more modular structure, which in turn provides for easier maintenance and testing. Current system features and operational details are discussed in an appendix.

Members of Advisory Committee: Dr. Vijayan K. Asari
Dr. Ravindra P. Joshi

To Mom, Dad, Bretta and Ben.
Thanks for all of your patience and support.

ACKNOWLEDGMENTS

I would like to thank Dr. Stephen Zahorian for the opportunity to work on this project and in the Speech Communication Lab, and for his infinite patience during the writing of this thesis.

I would also like to express my gratitude to those who donated their voices to the speech database, and especially to those helped with arrangements for speech database collection:

Ms. Sarah Balcom
Ms. Joyce Kiser
The Zahorian Family
Ms. Diedre Henriques
Mr. Thomas Hudgins
Ms. Arlene Ingram

Finally, I would like to acknowledge the National Science Foundation who partially supported this work under grant# BES 9977260

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF EQUATIONS

Equation Page

CHAPTER I

INTRODUCTION TO VATA

Purpose

This thesis describes the Vowel Articulation Training Aid (VATA), a speech therapy tool designed to supplement insufficient or missing auditory feedback for hearing impaired persons.* The purpose of the system is to provide a *visual speech display* (VSD) which gives on-screen feedback to a user about the quality of pronunciation of ten American-English vowel phonemes. Such feedback has been shown to be useful in speech training work with hearing impaired children (Correal, 1994). While the system is intended primarily as a speech therapy tool, it can be used in any situation where feedback about monophthong vowel pronunciation is desirable, such as in foreign language training.

VATA has been implemented completely in software as a program designed to run on a Win32 platform-based operating system such as Windows 95, 98, Me, NT, 2000 or XP. It requires only a PC-compatible computer with an installed sound card.

Description

The system has two main displays. One is a bargraph display, which gives feedback about how well speech utterances fit into discrete pre-defined phonetic (vowel, for the present implementation) categories. The other is an “ellipse” display, which provides continuously variable feedback about utterances. The system gives feedback for the sounds /ah/, /ee/, /ue/, /ae/, /ur/, /ih/, /eh/, /aw/,

* This thesis uses the APA style for citations, figures and tables.

/uh/, and /oo/, which correspond to the vowel sounds found in the words “cot,” “beet,” “boot,” “bag,” “bird,” “pig,” “bed,” “dog,” “cup,” and “book” respectively. The labels for this system (/ah/, etc.) were assigned by the ODU speech lab, and correspond to the ARPABET labels /aa/, /iy/, /uw/, /ae/, /er/, /ih/, /eh/, /ao/, /ah/, and /uh/ commonly used in speech processing literature. The labels assigned in the ODU speech lab were considered to be more similar to commonly observed spellings of these sounds in everyday words, and thus thought to be easier for small children to interpret.

The bargraph display (Figure 1) resembles a histogram, with one bar for each vowel sound of interest. The height of the vowel's bar varies in proportion to the accuracy of the speaker's pronunciation of that vowel. Correct pronunciation yields one steady, clearly defined bar, while the remaining bars assume zero or small values.

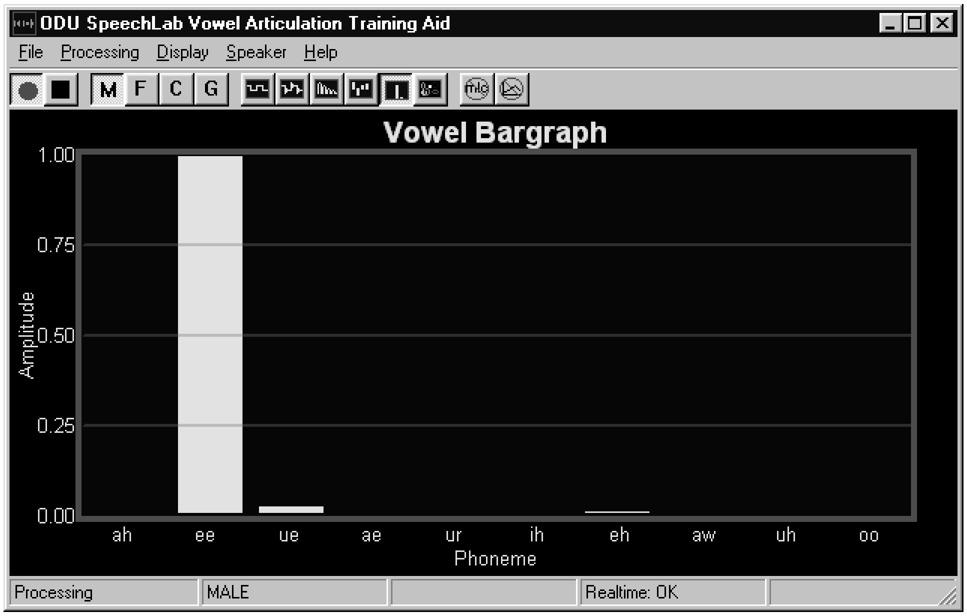


Figure 1. Bargraph display response for correct pronunciation of /ee/

The ellipse display (Figure 2) divides the screen into several elliptical regions, similar to a standard F1/F2 type display

(Peterson & Barney, 1952). Unlike the F1/F2 display, this 'ellipse' display bases its output on Neural Network Classifier output values obtained using Discrete Cosine Transform Coefficients (DCTCs) of the log magnitude spectrum rather than the formant frequencies of the utterance. Correct pronunciation of a particular sound places a basketball icon within the corresponding ellipse and causes the icon's color to match that of the ellipse. Incorrect or unclear pronunciation results in the ball icon 'wandering' about the screen or coming to rest in an area not enclosed by an ellipse. By observing the continuous motion of the ball, a speaker gains information about how to adjust his or her pronunciation in order to produce the desired vowel sound.

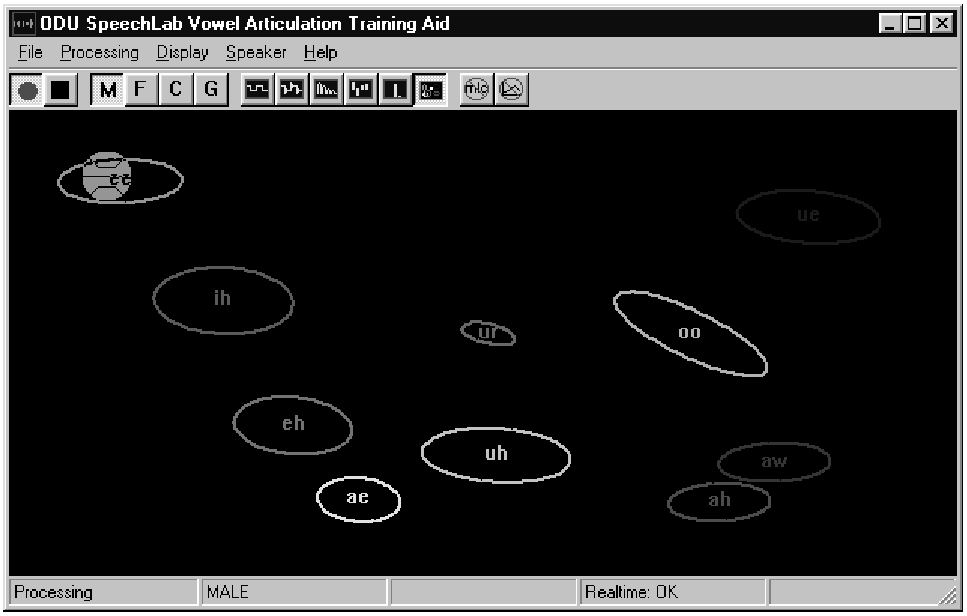


Figure 2. Ellipse display response for correct pronunciation of /ee/

A speaker group selection option with "CHILD," "FEMALE," and "MALE" settings allows both the ellipse and bar graph display modes to be fine-tuned for better classification of sounds produced by children, adult female, or adult male speakers respectively. A fourth speaker group option, "GENERAL" causes the display to attempt to classify sounds produced by a speaker from any of the three other categories.

Several other display modes show (in real time) the results at different points in the signal processing chain and are useful for system diagnostic testing and signal analysis. These other displays include the unprocessed acoustic time signal from the sound card, the filtered time signal, the signal's frequency spectrum, and the DCT coefficients (features) of the signal.

System Operation Overview

VATA has two modes of operation: The real-time speech-display mode, and an off-line training mode. In the real-time mode, the system accepts a continuous stream of audio data and produces the displays described above. No reconfiguration or adaptation of the neural network classifiers or any other elements of the signal processing chain occurs during real-time operation. All displays are produced using criteria established during the off-line training. The real-time mode also has several diagnostic functions available for gathering system performance data and verifying proper operation of the signal processing subsystem.

The second major mode, is the off-line, or "training" mode. When in this mode, support programs use pre-recorded examples ('tokens') of correctly pronounced phonemes to "teach" the real-time display system to recognize when a phoneme is correctly pronounced and to produce appropriate display feedback. The support programs handle various aspects of token data management and system training. The results of the system training are stored in several configuration data files for use by the real-time system.

In addition to the main executable programs, various support files are required to configure and manage the training process.

Scripts and "batch" files assist in data formatting and program execution sequencing, while "setup" files provide initialization data and configure options which moderate the training method process. Table 1 summarizes the major programs used in VATA's training and real-time modes. Appendix III provides a summary of all VATA system configuration files and their functions.

Table 1: Major programs comprising the VATA system

Program	Mode	Function
WinRec.exe	Off-line	Records training data tokens
WinSel.exe	Off-line	Selects tokens for training
TfrontC.exe	Off-line	Calculates token features
Scale.exe	Off-line	Scales the calculated features
Transfor.exe	Off-line	Performs data transformations
Neural.exe	Off-line	Trains neural network
Vata.exe	Real-time	Produces all real time displays

Overview of Processing Steps

A block diagram of the VATA signal processing steps is presented in Figure 3. The operating system interacts with the sound card to acquire a section of data (a “segment”) from the continuous audio data stream. Special driver software written for the VSD program, the “Wave API,” interacts with the Windows multimedia services to handle double buffering and ensure that no samples are lost. Since the driver software communicates with the operating system services and not directly the hardware, the VATA system should work with most commonly available Windows-compatible sound cards.

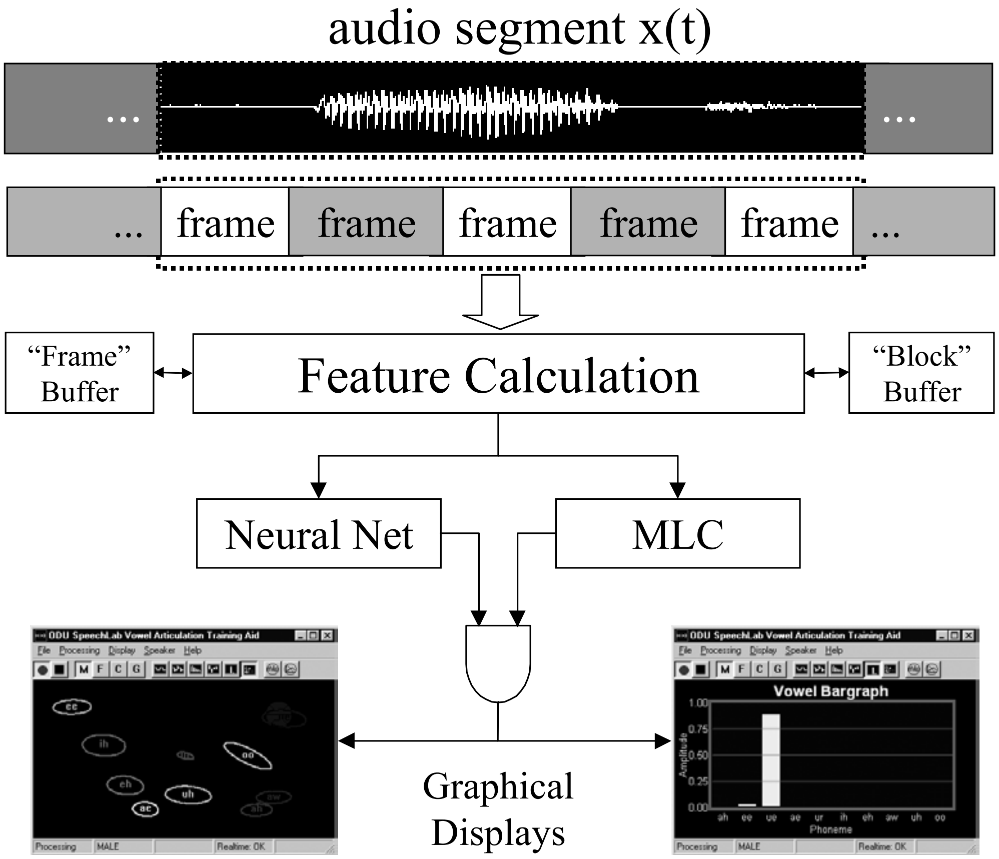

audio segment x(t)
...
...
... frame frame frame frame frame ...
"Frame" Buffer
Feature Calculation
"Block" Buffer
Neural Net
MLC
ODU SpeechLab Vowel Articulation Training Aid
File Processing Display Speaker Help
ee
ih
oo
eh
uh
ae
Processing MALE Realtime: OK
Graphical Displays
Vowel Bargraph
Amplitude
1.00
0.75
0.50
0.25
0.00
ah ee ue ae ur ih eh aw uh oo
Phoneme

Figure 3. VATA block diagram and signal processing steps

The first step in signal processing is to divide the audio segment into sub-segments called "frames." The frame is the basic unit on which signal processing is performed. Each frame is first passed through a high-frequency pre-emphasis filter, typically centered at 3.2kHz. Next a Fast-Fourier Transform (of typically 512 points) is performed on each frame and the base-10 logarithm of the magnitude of the resulting coefficients is taken. This "log-magnitude spectrum" is then averaged over several (usually about 5 to 10) frames, and a Discrete Cosine Transform (DCT) expansion (using the formula shown in Equation 1) is performed to yield the Discrete Cosine Transform ("Cepstral") coefficients.

$$DCTC(i) = \sum_{k=0}^{N-1} X(k)\phi(i,k)\ , \qquad \text{where} \quad \phi(i,k) = \cos\left(\frac{\pi i\left(k+0.5\right)}{N}\right)$$

Equation 1: Formula for the Discrete Cosine Transform Coefficients used in VATA

Other work has shown that modifying the basis vectors via a bilinear frequency warping function,

$$f' = f + \frac{1}{\pi}\tan^{-1}\left\{\frac{\alpha\sin(2\pi f)}{1\text{-}\alpha\cos(2\pi f)}\right\}$$

can substantially increase performance (Zahorian and Nossair, 1999). The typical value used for the bilinear warping coefficient α is 0.45.

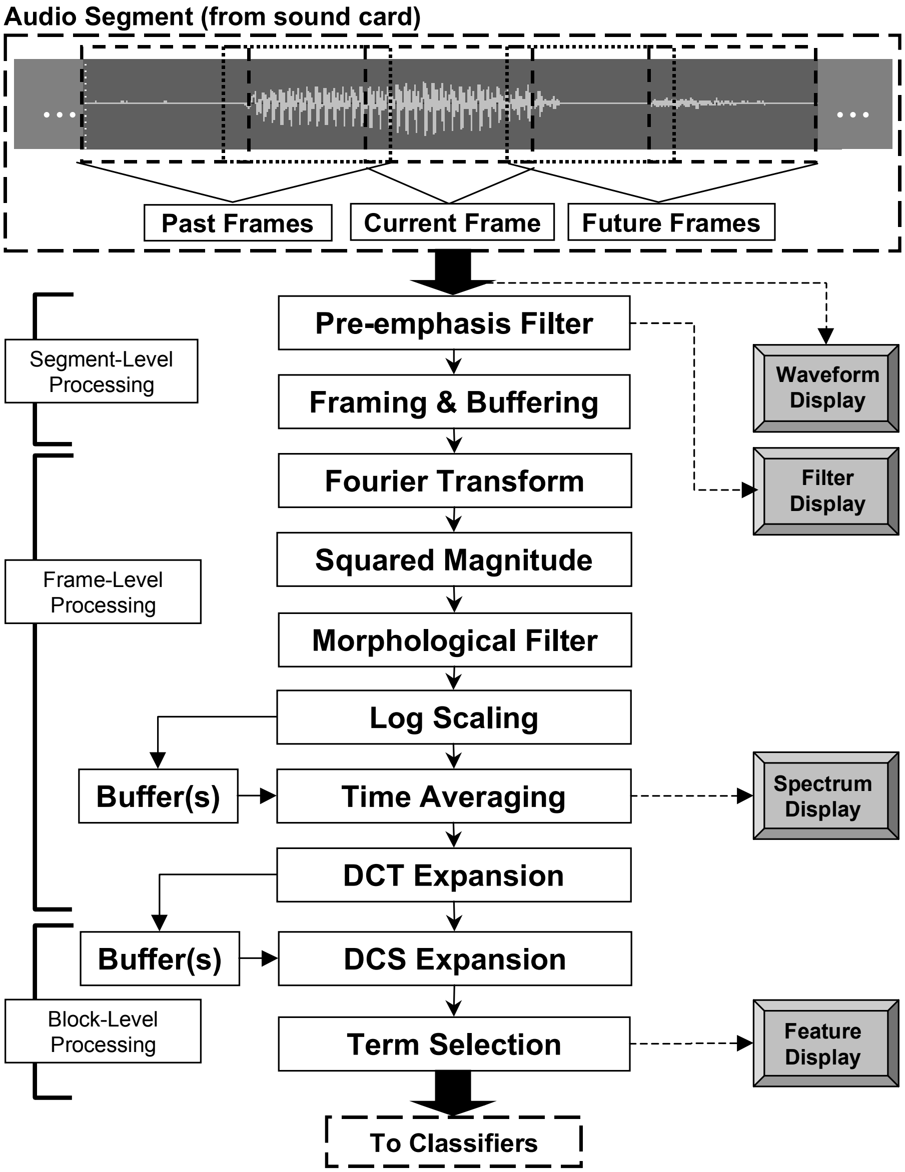


Figure 4: Feature calculation detail

Several frames' DCTC values can be combined to form a "block" of feature data over which a Discrete Cosine Series expansion can be performed to encode feature variations over time. This step is useful for encoding feature changes in the analysis of non-steady-state sounds such as diphthongs and some consonants, but is not typically performed in the analysis of monophthong vowel sounds. For fairly stationary sounds such as monophthong vowels, it can be assumed, in theory at least, that the features do not vary significantly over time. In practice, it is shown that there will be some feature variation due to windowing effects and system noise (these effects are considered in Chapter III, "Noise and Feature Stability").

The first several (typically 12) cepstral coefficients (DCTCs) are then normalized (zero mean and standard deviation of ±.2, or typically a range of ±1) and passed to a neural network (NN) classifier and a maximum likelihood classifier (MLC). The NN, a feed-forward multi-layer perceptron, attempts to classify the speaker's utterance into one of the ten vowel categories. The neural networks used in VATA have one input node per feature, 15 to 25 nodes in the hidden layer, and ten output nodes (one per vowel class). The NN's are trained using error backpropagation for (typically) 250,000 iterations.

The MLC block (developed as part of this thesis work, and described in more detail in chapter IV) then verifies the NN's classification of the current signal by calculating the distance between the current features and the known distribution of features for the vowel reported by the NN. If the current signal's features are consistent with the statistics of the vowel reported by the NN, the display produces feedback that signifies correct pronunciation. If not, it does not report that a “proper” vowel has been recognized. In this case the Bar Graph will set all vowel categories to zero, and the ellipse display will not change the color of the ball icon to match any ellipse in which the ball lands. For purposes of subjective system performance evaluation, the MLC processing can be disabled by means of a program switch.

Data Management and Real Time Operational Considerations

The basic units used for data management can be described as follows. The “segment” interval is used to double buffer the data for exchanges between the sound card, the data processing, and graphics displays. That is, the continuous incoming speech signal is divided into segments (typically of length about 100 ms). Whenever the A-to-D converter reports that a segment buffer is full, then the following actions must occur:

1. The two segment buffers are interchanged, so that the next section of the signal is placed in the other segment buffer. Note that this action is handled by the wave API module and therefore does not need to be managed directly by the VATA application.

2. The final results of processing the data from the previous segment are displayed on the graphics screen. Thus the graphic display is updated once per segment interval time.

3. Processing of data in the just filled segment must begin, and must be finished before the next segment is filled, in order to sustain real-time operation.

During the processing of each sampled audio “segment,” the software analyzes sections of the input signal that equal to the “frame length” interval. This interval, typically about 20 to 30 ms, is commonly used in speech analysis. Note that the program also uses a “frame space” time, which can be less than, equal to, or greater than (but which is typically less than) the frame length, and which controls the rate at which spectral analysis is repeated. The software has been written so that the frame time intervals are independent of the segment intervals. For example a frame can span an interval of speech signal contained in two segments.

The third data management unit is the “block length.” A “block” is a time interval corresponding to an integer number of frames. Features and neural network computations are done once for each block. There is also a block spacing parameter (which again can be less than, equal to, or greater than) the block length, which controls the rate at which block calculations are repeated. Note that, in principle, block time and spacing are also independent of the segment interval. However, in practice, the block time and block spacing are usually set so that one and only one new block of features is computed for each segment, since the display is updated only once for each new segment. Thus typically in the VATA system, segment time, block length, and block spacing, are all set to equal about 100 ms.

Thesis Goals and Summary

The overall goal of this thesis was to continue work for the further development of the ongoing VATA project. The specific objectives were:

1. To increase the size of the training database for VATA.

2. To investigate in detail sources of spectral variability for vowels, especially focusing on the use of sound cards as a possible noise source.

3. To develop and implement the MLC classifier addition to the neural network, to improve out of category rejection.

4. To examine vowel classification accuracy as a function of amount of training data, neural network size, frequency range, and sampling rate.

5. To modify the VATA software to be more useful for future research, and also to incorporate the results found from objectives 1-4 to improve vowel recognition performance.

The process for assembling a large database of speech recordings is documented in Chapter II. Examination of noise and feature stability issues and are explained and presented in Chapter III. Chapter IV explains the Maximum Likelihood Classifier which was incorporated to improve the VATA's usefulness by reducing the number of "false correct" feedback indications that it gives. Results of an in-depth objective performance evaluation of the VATA system are presented and discussed in Chapter V. Conclusions drawn from this work are presented in Chapter VI.

Three appendices are also included. Appendix I provides documentation the VATA real-time program structure and features included in the current version. Appendix II presents a full description of the speech recording database assembled during the course of the work presented in this thesis. Appendix III provides a complete listing of the VATA system configuration files and the function performed by each.

CHAPTER II

SPEECH DATABASE COLLECTION

Introduction

For “good” performance, a neural network classifier requires a large amount of training data. That is, for a typical neural network classifier, the test recognition rate (the number of correct classifications given a set of unknown feature vectors) improves as the amount of training data the classifier receives increases (Figure 5). Previous work with smaller speech databases has shown that the VATA neural network classifier follows this trend (Figure 6) (Zimmer, Dai, and Zahorian, 1998). Note that it is the recognition rate on test data, not training data, that is a good indication of the expected performance of the classifier in actual operation. The effects of the size of the training database on both training and test recognition results when a significantly larger amount of speech data is available are explored in more detail in Chapter V.

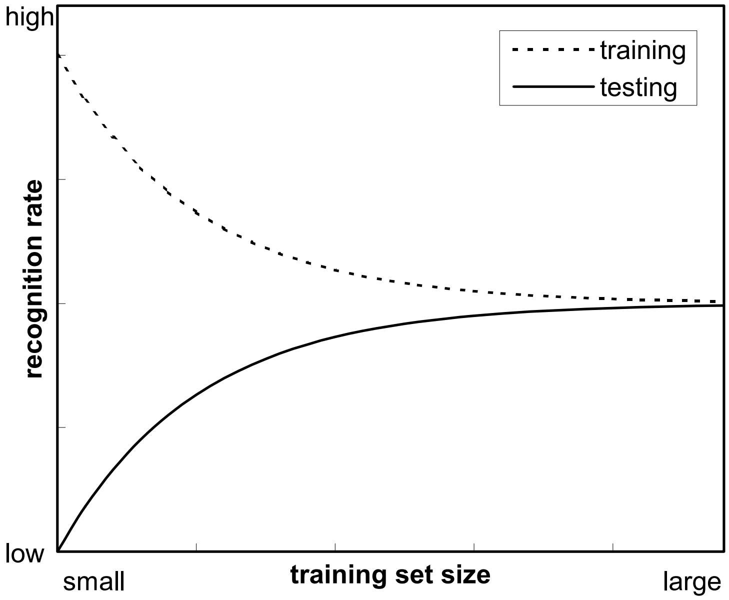


Figure 5. Expected neural network recognition rate trend as a function of training data set size

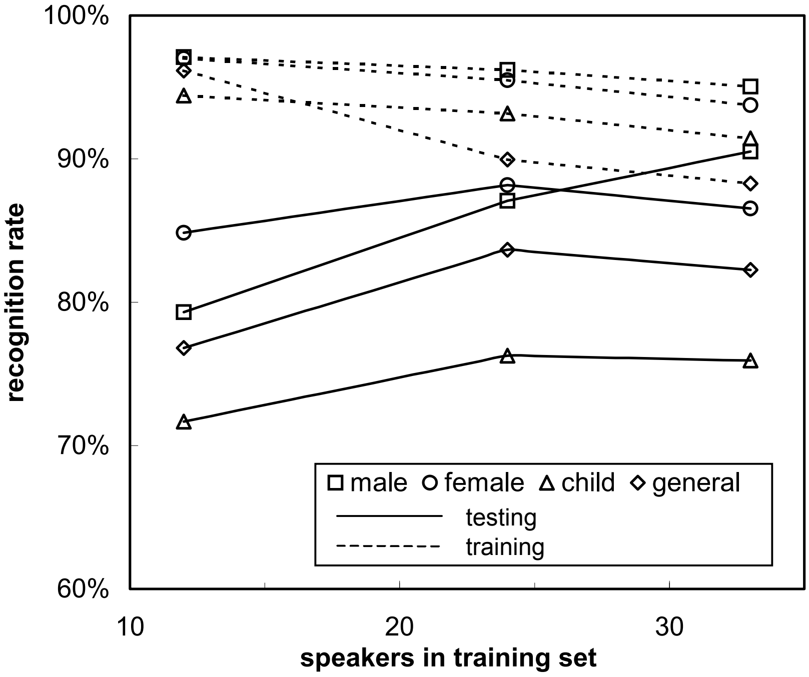


Figure 6. Observed recognition rate behavior as a function of training set size for a small training database

Thus, one goal of this thesis was the collection of a large recorded-speech database for use as training and testing data for neural network classifiers. The collected database has three separate recordings of each of 10 monophthong vowels, three diphthong vowels, and thirteen consonant-vowel-consonant (CVC) word tokens for every speaker (78 recordings per speaker).

The database contains three speaker categories: Child (both male and female), Adult Female, and Adult Male. Child speakers were generally selected to be between five and thirteen years old, and adults were generally selected to be over 13 years old. In the case of adolescents who were between ages 11 and 15 the speaker category was chosen based on the vocal quality of the subject; that is, those child subjects whose voices had matured were included with adult speakers. If this distinction was not clear, the cutoff age was set arbitrarily at 14 years. Table 2 summarizes the data collected over the course of this thesis in terms of number of speakers by token category and speaker category contents.

Note that all speakers produced the vowel sounds, whereas not all speakers produced the CVCs. Thus the total number of speakers in the database was 314. The total number of vowel tokens collected was 12,841, and the total number of CVCs collected was 11,500. To date, and especially for the work reported in this thesis, only the vowel tokens have been used. For the initial recordings (the first 195 speakers), a sampling rate of 11.025 kHz was used. The remaining speakers (119) were recorded using a sampling rate of 22.050kHz (Table 2). More discussion of the effect of the change in sampling rate is given in chapter V.

Table 2: Breakdown of speech database totals by speaker group and recording type

	Speakers				Recordings		
Database	**Male**	**Female**	**Child**	**Total**	**Vowels**	**CVCs**	**Tokens**
V_0	1	5	13	19	1381	5	1386
V_CVC_0	18	11	21	50	1917	1930	3847
V_CVC_1	55	47	24	126	4902	4868	9770
V_CVC_2	40	75	4	119	4641	4702	9343
Total	114	138	62	314	12841	11505	24346
Token Type							
Vowels	114	138	62	314	12841	11505	24346
CVCs	113	133	49	295	11460	11500	22960
Sampling Rate							
fs=11050 Hz	74	63	58	195	8200	6803	15003
fs=22050 Hz	40	75	4	119	4641	4702	9343
Total	114	138	62	314	12841	11505	24346

Recordings were collected using software developed in the speech communication laboratory for "multimedia" computers. The software provides automatic prompting, recording, signal endpoint detection and labeling, and cataloging facilities. This software helped to reduce the work involved in organizing the thousands of recordings and information files comprising the complete database.

Description of recording equipment

The "multimedia" computers used in the speech database collection consisted of Intel 80x86 based computers running the Windows 95 or NT operating systems. All computers used either built-in or commonly available "sound cards" for analog to digital conversions of the speech signals. Table 3 summarizes the sound cards used.

Table 3: Computers and sound cards used to record speech samples for database

Computer	Sound Card
Desktop 1 Windows NT 3.51	Creative Labs Sound Blaster 16
Desktop 2 Windows NT 4.0	Creative Labs Sound Blaster 16
Desktop 3 Windows 95	Creative Labs Sound Blaster 16 Value PnP
Desktop 4 Windows NT 4.0	Creative Labs Sound Blaster AWE64
Desktop 5 Windows 95	Turtle Beach Multisound "Classic"
Laptop 1 Windows 95	Compaq Armada 1520 Built in sound card (ESS Audio Drive 1875)
Laptop 2 Windows 95	New Media Labs WAVJammer PCMCIA Audio Card

Most recordings were made with inexpensive, high-impedance dynamic microphones, typical of those included with a multimedia computer.* The microphone was usually fitted with a foam windscreen to reduce breath noise and to reduce large "breath pop" transients typical of the plosive /b/ and /p/ sounds. Additionally, to eliminate microphone handling noise and to facilitate consistent microphone placement, the microphone was mounted on an adjustable boom stand, or in some instances, a desk stand.

The majority of the recordings were made in the ODU Speech Communication Lab. Those recordings not made at the Speech Lab were made "on location" at a variety of sites, including other locations on the ODU campus, area schools and private residences. In each case, the recordings were made in as quiet a location as was possible. Each complete recording session lasted about 15 minutes. Recording subjects have included many Old Dominion University students and staff, students

* Approximately 10 speakers were recorded using a head-mounted microphone. It is planned that future recordings and work with VATA will use headset microphones.

and staff of area primary and secondary schools, as well as family and friends of persons working in the ODU Speech Lab.

Data Review and Evaluation

During speaker recording sessions, efforts were made to ensure that only properly articulated, high-quality tokens are accepted. Nonetheless, review of some of the recordings in the database showed that several of the tokens contained spurious vocalizations, breath noise, or in some cases, the wrong sound or word (e.g., "beet" instead of "dog"). To prevent these "bad" tokens from introducing error into further experiments, a complete review of all recorded data was undertaken.

Each of the recorded tokens for a particular subject was played and evaluated. Tokens were examined for quality of articulation (i.e., acceptability as a "good" example of the vowel or word that it represents) and recording quality (i.e., freedom extraneous noise and undesired vocalizations, and consistency with other recordings with respect to amplitude). Tokens that did not meet the acceptance criteria were tagged as unusable, thus removing them from the training and testing data.

Token quality was judged first by examination of the time-amplitude waveform. Tokens that had a large amount of background noise and/or large non-speech transients were tagged. Second, the quality of a token's articulation was judged (subjectively) by the author, using a simple "reasonable listener" criterion-- i.e., could a reasonable native speaker of American English listening to the token recognize it as the token that it purports to represent? If the answer was yes, the token was of acceptable quality, and left in the database. Tokens judged to be of poor quality were also marked as unusable.

Another important variable in the collected speech data is the signal endpoints. Care must be taken to ensure that signal endpoints are detected consistently. While the data collection software attempts to automatically detect endpoints, it is limited in its ability to discriminate between speech signal and other spurious sounds, which may occur during the recordings. This can often result in mislabeled endpoints.

Initially steps were taken to re-label endpoints by hand, but preliminary trials suggested that this would likely reduce overall system performance. Endpoint labels applied using automatic techniques appeared to yield better results. The method used to automatically label endpoints is based on a method described for endpoint detection in a noisy office environment (Evangelos, Fakotakis, & Kokkinakis, 1991)

While no carefully controlled experiments have been performed to evaluate any effects that "poor token" removal might have on NN recognition rates in this system, prudence suggests that training data used to set up the VATA system should be of as high quality as is possible.

CHAPTER III

NOISE AND FEATURE STABILITY

Background

Comparison of the newer Windows-based vowel speech displays with a similar, previous system developed using a TMS320C floating-point DSP platform and custom-built preamplification circuitry showed that while the newer display was usable, it appeared to be considerably less stable than was the older system’s display. The instability was most noticeable in the FFT and feature displays. These spectral and feature effects were the subject of experiments conducted by Bingjun Dai, and reported in his Master's Thesis (Dai, 1998). Mr. Dai reported that regardless of the input signal (either a trained speaker producing a steady vowel sound into a microphone or direct sinusoidal input from a signal generator) the FFT and feature displays would never settle to a steady state, but would randomly fluctuate. When the same signals were applied to the older system’s inputs, a steady display could be observed. Unfortunately, the experiments did not examine the problem in depth and yielded no insight into the cause of the new system's instability.

The signal processing algorithms used in the Windows-based system were tested in MATLAB for proper operation, and appeared to be working correctly. The feature instability was assumed to be attributable to one of three possible sources: (1) noise in the signal acquisition system, (2) program code errors which could have introduced discontinuities in the input signal, or (3) fundamental variability in the speech signal, which was simply not observed in the older system. The objective of this chapter is to investigate these three possibilities, with particular emphasis on the inexpensive sound cards used in the new system in place of the low-noise A/D converters used on the dedicated DSP card of the previous system.

Sound Cards as a Possible Source of Noise

Detailed specifications for the multimedia sound cards were often not directly available in the product documentation. Manufacturer's specification reported in product documentation were sometimes incomplete, not describing the conditions under which the test measurements were taken or how the measurements were made. Contacting the manufacturer's technical support staff often yielded no additional information. Reviews of the equipment by various audio and computer magazines were somewhat more informative, but the reviews did not provide sufficient documentation of their test techniques or results.

Time constraints precluded an exhaustive review of the available sound card signal-to-noise ratios. Previous work had suggested that the underlying signal processing algorithms and program code were correct and reliable (Dai, 1998). To address the display instability problem, a set of simple tests was devised to assess spectral and feature stability for different sound cards. As a control for the sound card testing, several tests were also made of spectral and feature stability, with controlled signals inserted in the signal processing path, bypassing the sound card.

In order to investigate the spectral and feature stability issues, experimental tests were conducted as follows:

1. Comparison of various sound cards in terms of background noise and frequency response.

2. Spectral and feature stability using a square wave input signal for the various sound cards, as well as a control version not using the sound card.

3. Spectral and feature stability using prerecorded digitized vowel sounds, again using the various sound cards for signal acquisition.

The procedures for these tests, and results, are given in the remainder of this chapter.

Test Signals

In order to make these tests, a CD was created with several test signals. These signals were all recorded using the standard CD-Audio sampling rate (44.1 kHz). Speech recordings were made using an ElectroVoice ND257A Microphone, Mackie 1604VLZ mixer microphone preamp, and an Echo Audio “Layla” 20-bit professional grade sound card, and Recorded using Syntrillium Software’s “Cool Edit 96.” The test signals also were created with Cool Edit using the program’s built-in signal generation functions. All signals were stored using Resource Interchange File Format (RIFF) "wave" files. The test signals included:

1. Sine wave signals, from 31.5 Hz up to 8 kHz, with frequency spacing in powers of 2 (8 different frequencies), with maximum signal amplitude set to –6dB from maximum deflection.

2. Square wave signals, same format as above, except maximum fundamental frequency generated was 1 kHz (5 total signals)

3. Frequency sweeps, beginning with a 1 second 1 kHz test tone, followed by .5 seconds of silence, followed by a sine wave sweeping from 1 Hz to f_{max} over a 4 second interval, with maximum signal

amplitude of –3dB from maximum deflection. The f_{max} values used were 5512, 11025 and 22050 Hz (3 total Signals).

4. White noise, with an rms amplitude 6 dB below the dynamic range allowed, 60 seconds long.

5. Vowel signals. One token for each of the ten vowels used in the VATA (as mentioned above) for one male and one female speaker, each attempting to hold each vowel sound as steady as possible for 10 seconds. These were recorded with the sound system mentioned above at 44 kHz. Thus there were a total of 20 recorded vowel tokens.

Note that not all of these signals were used for the formal tests reported in this chapter. Based on informal pilot testing, the key points of concern were addressed using only a small number of the total signals. In particular, the three vowels (/ah/, /ee/, /ue/) were judged to be adequate to demonstrate spectral and feature stability for vowel sounds.*

For use in direct-signal insertion testing of the VATA system, each "master" 44 kHz signal was downsampled to 22 and 11 kHz. The downsampling was done using Cool Edit, which uses a lowpass digital filter prior to each pass of downsampling. Signals were assembled to a Mixed-Mode CD, allowing the master (44.1 kHz sampled) test signals to be reproduced using a standard CD-Audio player, and the original and downsampled versions of the recordings to be stored in the data track of the CD, readable using a standard CD-ROM drive.

* The vowel sounds mentioned here and used for the more formal testing were chosen since these vowels are widely separated and are considered to best span the "vowel space." Results are only given here a subset of the complete range of sounds created since pilot testing indicated that these signals appeared to be sufficient to illustrate the properties under investigation.

Sound Card Testing

The sound cards listed in Table 4 were selected for examination (The computer used for the desktop sound card testing was the same "Desktop 4" listed in Table 3 used for speech database recording, while "Laptop 3" was a new computer different from the laptops used in database recording). The soundcards were separated into three categories: Low, Average and High "quality." The IBM Thinkpad (Model 390x) built-in sound card was selected to represent "low-quality" sound cards, the Creative Labs Ensoniq PCI and SoundBlaster 16 to represent "average-quality" sound cards, and the Creative Labs PCI512 to represent "high-quality" soundcards. Note that in this context "quality" is a subjective measure based on cost and manufacturer claims.

Table 4: Computers and sound cards used for sound card testing

Computer	Sound Card
Desktop 4 Windows NT 4.0	Creative Labs Ensoniq PCI
Desktop 4 Windows NT 4.0	Creative Labs Sound Blaster 16
Desktop 4 Windows NT 4.0	Creative Labs Sound Blaster 512 PCI
Laptop 3 Windows NT 4.0	IBM ThinkPad 390x (Stock IBM Sound Card)

To facilitate repeatable control testing the VATA real time display code was altered to provide several new features:

1. A built-in test tone generator, capable of synthesizing sine, square, sawtooth, and triangle waveforms with variable period and

amplitude, all of which could be inserted directly into the signal processing path.

2. Capture of real-time display data for the "frequency spectrum" and "feature" (DCTC) displays to a file for off-line analysis by other programs such as MATLAB

3. Real-time display operation using a pre-recorded audio file as the audio source.

4. Capture of audio data from the sound card to an audio file.

5. Other miscellaneous command-line options to facilitate real-time system testing in a 'batch' mode.

Once the VATA code had been modified to facilitate testing, three sets of experiments were devised to test the soundcards performance in the VATA system. The following paragraphs describe the experiments.

Sound Card Frequency Response and Noise Tests

The goal of the first experiment was to compare the noise floor and frequency response measurements for each of the sound cards. Measurements were performed in a soundproofed room using the setup shown in Figure 7.

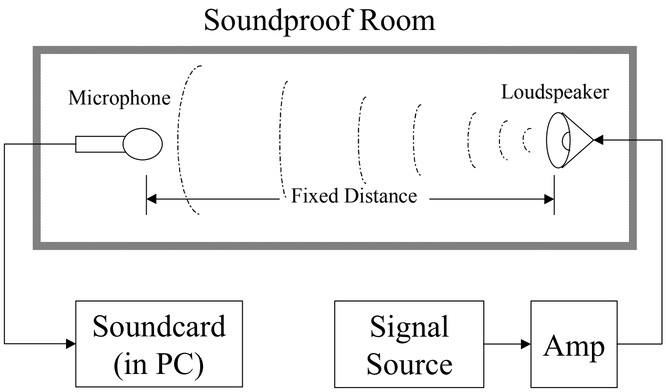


Figure 7: Setup for sound card tests

The microphone used for the soundcard tests was a unidirectional dynamic (impedance 600 ohm) DM-20SL (shipped with a commercially available speech recognition package) the loudspeaker was an Altec-Lansing Model FM40526511 (a typical computer speaker supplied with "multimedia" personal computers). The distance between loudspeaker and microphone was about 2 cm, and was held constant for all tests.

GAIN NORMALIZATION

Procedure

The first step in the procedure for measuring noise and frequency response of the various sound cards was to normalize the gain settings for microphone preamps between the various sound cards. A signal source (CD player playing tracks from the previously described test signal CD) provided a 250Hz square-wave signal that was reproduced by the amplifier and loudspeaker. The amplifier gain was set such that the SPL measured at the microphone measured 100 dB (C weighting). Once the basic test signal level was established, the gain of the soundcard's microphone input was adjusted (with all of the sound card's other input gains muted or deselected) such that the input signal meter provided in

Cool Edit registered -1.5dB from full deflection. This procedure was repeated at the start of testing for each of the different soundcards.

A similar procedure was employed for tests made using the auxiliary inputs. In this case, the line output of a CD-Audio player was connected to the auxiliary input and the same 250Hz square wave was played. With the sound card's other input sources muted or deselected, the input gain of the sound card's line input was adjusted again to produce –1.5dB deflection on Cool Edit's signal level meter.

Note that gain normalization among the different sound cards revealed that the microphone inputs of the SoundBlaster and Ensoniq cards apparently featured some form of Automatic Gain Control (AGC) that boosted microphone input levels dramatically. These cards required that the software gain controls be set to about 2/3 of the maximum setting to provide sufficient gain to record the 100 dB (SPL) test signal at -1.5dB from maximum amplitude. The IBM ThinkPad apparently had substantially less gain at the microphone input, apparently due to lack of such an AGC function, and yielded a signal amplitude of –20.5dB from maximum deflection at the maximum gain setting. None of the cards' documentation made mention of the presence or exact specification of the microphone input gain stages. Gain levels for the auxiliary (line) inputs of the different cards did not show great differences, typically requiring a setting of between 1/2 to 2/3 to achieve the target signal level.

NOISE LEVEL MEASUREMENT

Procedure

Once the microphone preamp gain for a soundcard had been set, Background noise level measurements were taken. Two sets of recordings were made for each soundcard: one with the microphone switched ON, to record the background noise of the loudspeaker/microphone/soundproof room system, and a second with the microphone switched OFF, to establish the level of background noise produced solely by the soundcard itself.

A third set of recordings was made with the microphone switched ON in the laboratory to capture the typical background noise level of the VATA system's operating/recording environment. The first 50 segments of each of these recordings were then processed by the VATA system, using the option to capture the FFT and feature values to data files, and data were analyzed using MATLAB.

Results

Figures 8,9, and 10 depict the noise spectra for each card for various conditions and locations. Note that the peak in each response at about 3.2 kHz is due to the peak in the pre-emphasis filter of the VATA processing code at this frequency. Using Figure 8 as baseline measure of "intrinsic" sound card noise performance (i.e., with mic off, thus implying no acoustic noise), the noise is generally highest for the IBM ThinkPad, lowest for the SoundBlaster PCI 512, and intermediate for the other two cards. However, the differences are not large -- approximately 13 dB between the best and worst card, as measured in the 3.2 kHz region. For the microphone on in soundproof room (Figure 9), the peak noise level for three of the cards is approximately the same, and only about 4 dB higher for the IBM ThinkPad. For the mic on in the lab condition (Figure 10), the noise level is actually the lowest for the IBM ThinkPad, and highest for the Ensoniq and SoundBlaster 512 PCI, and intermediate for the SoundBlaster 16. However, as alluded to previously, the unknown characteristics of the AGC makes it difficult to compare results among the soundcards for the conditions tested in Figures 9 and 10.

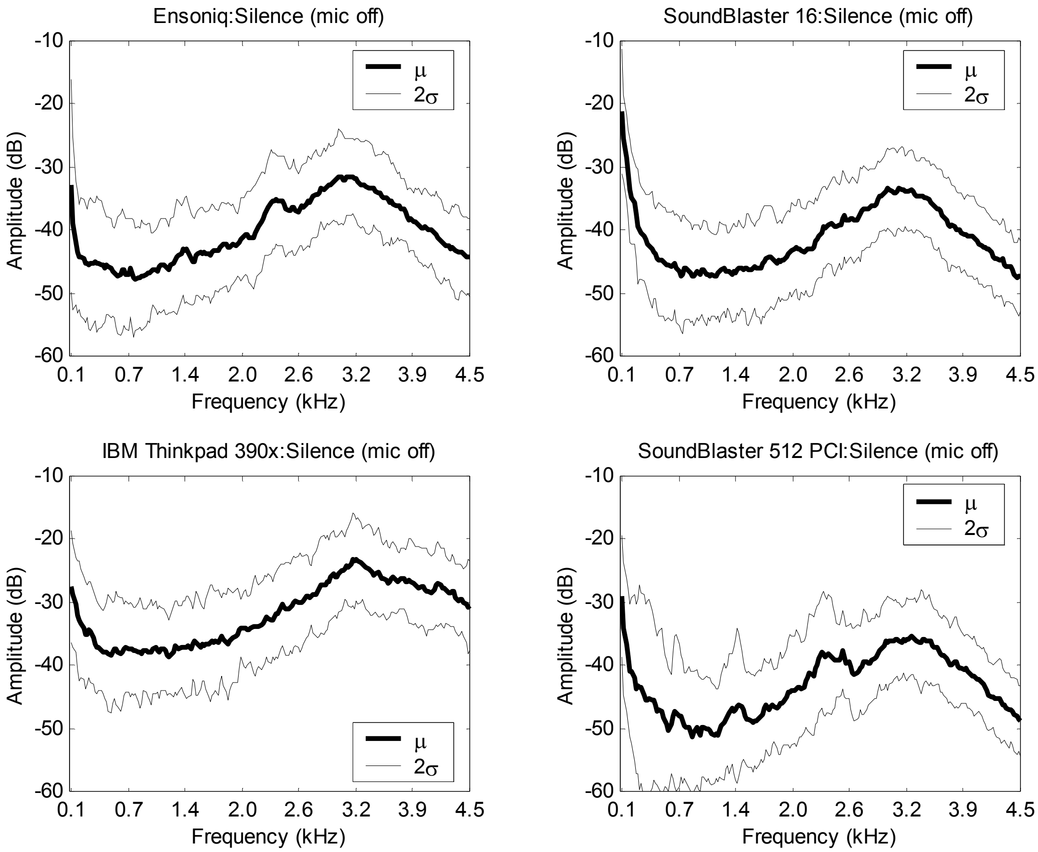


Figure 8: Average noise spectra of soundcards, no input signal.

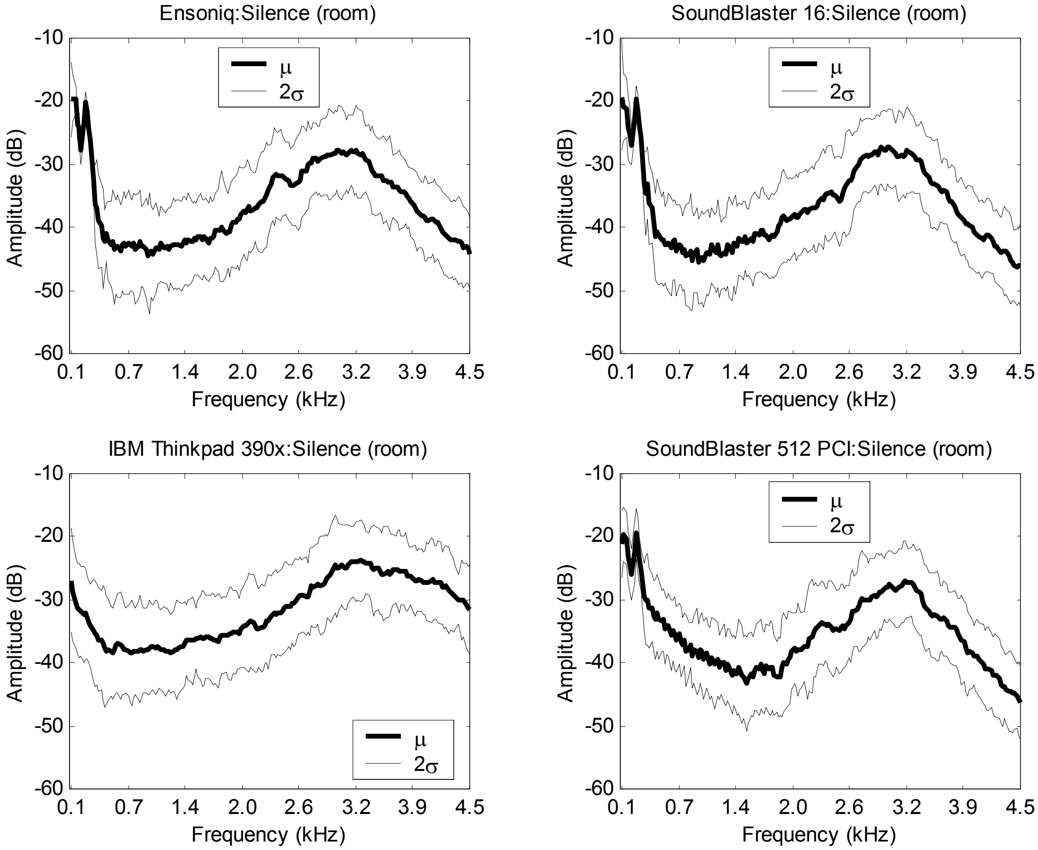


Figure 9: Average frequency spectra of sound cards in soundproof room, microphone on.

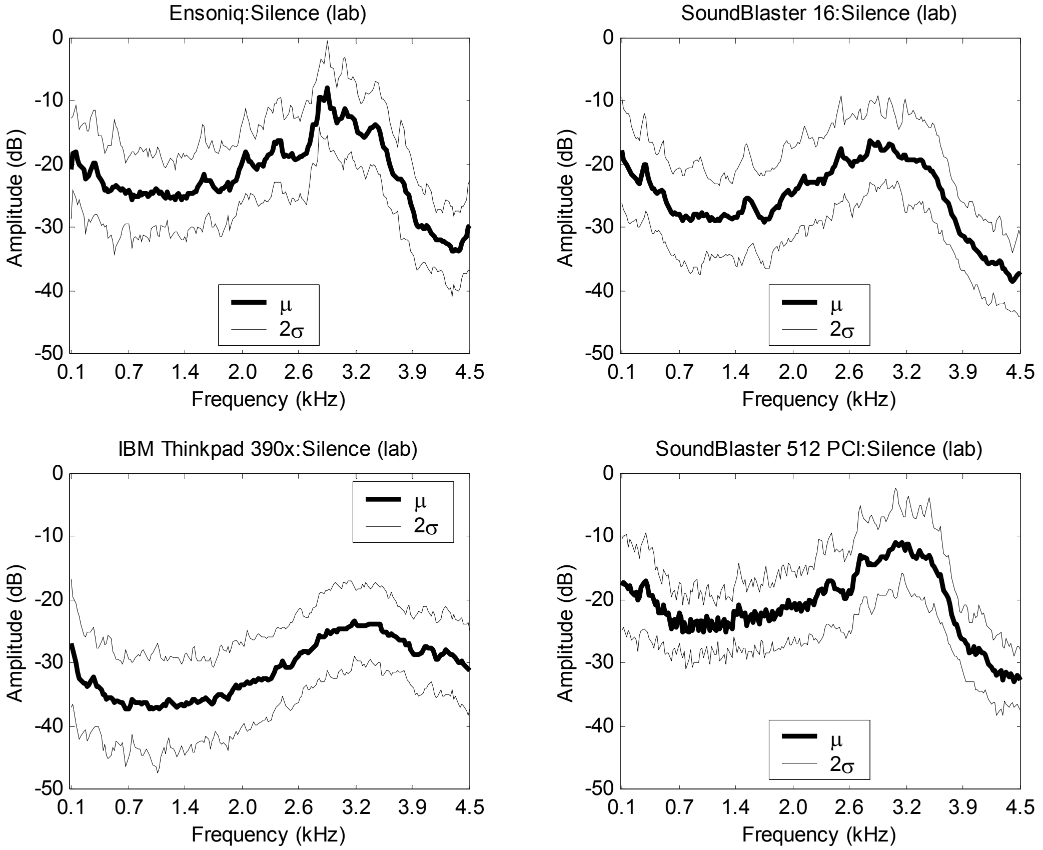


Figure 10: Average frequency spectra of sound cards in laboratory, microphone on.

FREQUENCY RESPONSE MEASUREMENT

Procedure

Once the microphone preamp gain had been normalized and preliminary noise measurements were completed, the next step was to measure each card's frequency response. Two measurement techniques were used. First, to measure average frequency response, a white noise signal was recorded for 60 seconds. The FFT values for each frequency bin were then averaged over the entire interval to estimate the frequency response. The second frequency response measurement was performed using a timed sine-wave sweep signal. A test signal was produced which had a 1s 1kHz signal tone followed by 0.5s of silence and finally a sine-wave linear frequency sweep over 4s which rose from 1 to 5512 Hz. The test signal was reproduced using the same amplifier and speaker, and recorded by each sound card using the same microphone and previously determined normalized gain settings. The recordings were then truncated to contain only the samples recorded 1.5s after the onset of the 1kHz signal tone to 5.5s after the onset of the 1kHz signal tone. Measuring the RMS amplitude of the recorded signal at specific times in the recording when frequencies falling in a specific frequency bin are present yields a separate measurement of the sound card's frequency response.

Results

Tt was observed that with the signal inserted at the line inputs, bypassing the microphone preamps, the sound cards all had reasonably flat (± ~3dB over the frequency range 50-3kHz) frequency responses (Figure 11). The SoundBlaster 16 and Ensoniq cards had nearly identical response graphs, with a gentle roll-off beginning at around 2kHz. The SoundBlaster PCI512 and IBM ThinkPad 390x cards each showed greater variability across the same range. The PCI 512 exhibits rounded peaks at 1kHz and 4kHz, with a fairly dramatic roll-off as frequency approaches fs/2 and the IBM ThinkPad 390x shows a single pronounced peak at 4.5kHz followed by a sloping roll-off.

Figure 12, which depicts the results of measurements taken using the speaker/microphone/preamp system, shows that while the frequency response of the system is affected dramatically by the inclusion of the additional components, the relative trends of the frequency responses are similar to those indicated in Figure 11. Thus the change in the frequency response graphs, while large, is generally attributable to the system transfer function of the speaker/microphone system. Note that frequency response results obtained with the frequency sweep method, are not shown here,since they were quite similar to the results obtained with the white noise test signal.

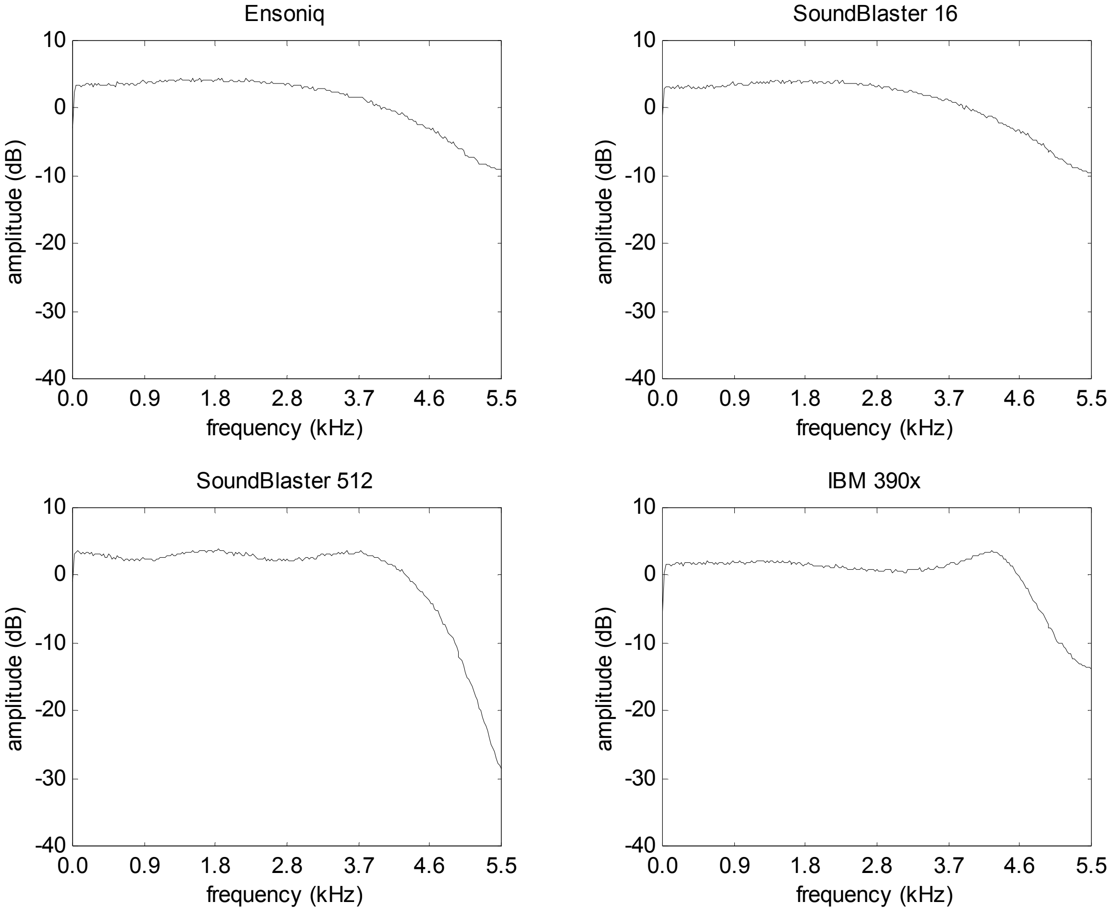


Figure 11: Frequency response of sound cards to white noise signal (line input).

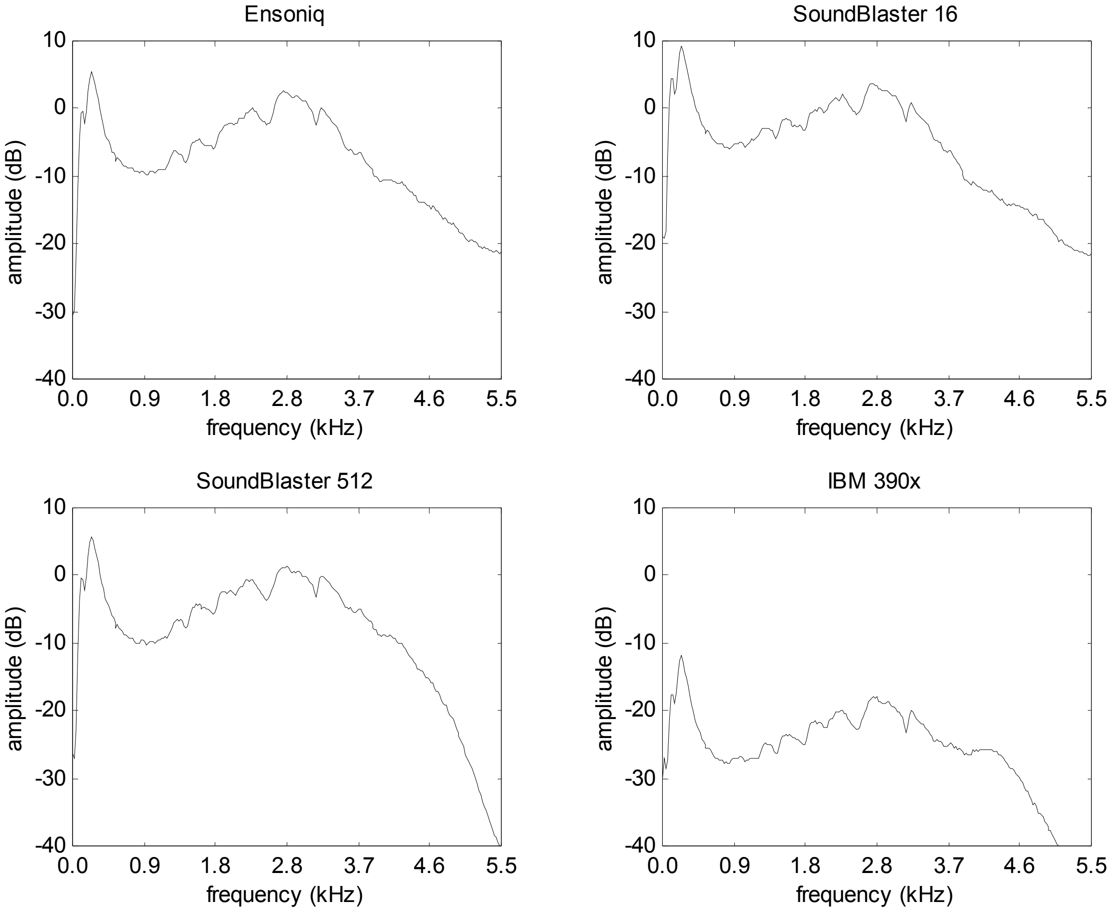


Figure 12: Frequency response of sound cards to white noise signal (mic input)

Spectral and Feature Stability Tests with test square waves

The next set of tests was designed to systematically investigate spectral and feature stability with known test signals--in particular a square wave with a 250 Hz fundamental. Although the main goal was to examine the sound card effects, as a control, tests were first done using only the digitized square waves inserted directly into the signal processing to examine “window” effects, using some of the features of the program mentioned previously.

To examine these “window” effects, two control tests were done. In the first case the sampling rate, fundamental frequency, and frame spacing were all adjusted so that there would be an integer number of test signal cycles in each frame, resulting in an identical signal from frame to frame. Thus, in theory, the spectrum and features computed for each frame should be identical. The parameters used for this test were a fundamental of 250Hz, a sampling rate of 10000Hz, segment length of 100ms and a frame length of 32ms.

For the second control test, the frame rate, fundamental frequency, and sampling rate were chosen as typical values used in VATA operation (fundamental=250Hz, f_s=11025Hz, segment=100ms, frame=30ms) with no attempt to maintain synchronization of the frames with the signal. Thus the windowing effect would be expected to result in some differences in spectral analysis from frame to frame since the signal would be different from frame to frame.

The tests with the sound cards used the prerecorded square waves, played through a speaker, and recorded by a microphone as described in the previous section. The parameters were the same as in the second control test. These tests were then done with all four sound cards mentioned. However, since the results from the previous section indicated only relatively small differences among the top three sound cards with respect to frequency response and noise characteristics, from this point on results are only given for the "best" (SoundBlaster 512 PCI) and "worst" (IBM ThinkPad) cards tested.[*]

For all control tests and sound card tests in this experiment, the spectrum and features were recorded for 5 seconds (50 segments) in a file. The mean and standard deviation of the spectrum and features were then computed.

[*] The plots for the two intermediate quality sound cards were inspected, for the square wave test signal and the recorded vowel sounds mentioned. Both were generally very similar to each other and closely resembled those obtained with the SoundBlaster 512 PCI.

RESULTS

The results for the tests mentioned above are given in Figure 13 (test signal directly inserted into signal processing chain) and Figure 14 (test signal recorded with sound cards). The figure panels give both spectral plots and feature plots, indicating mean values and standard deviations for each case. These plots thus give a measure of spectral and feature stability for the two cases.

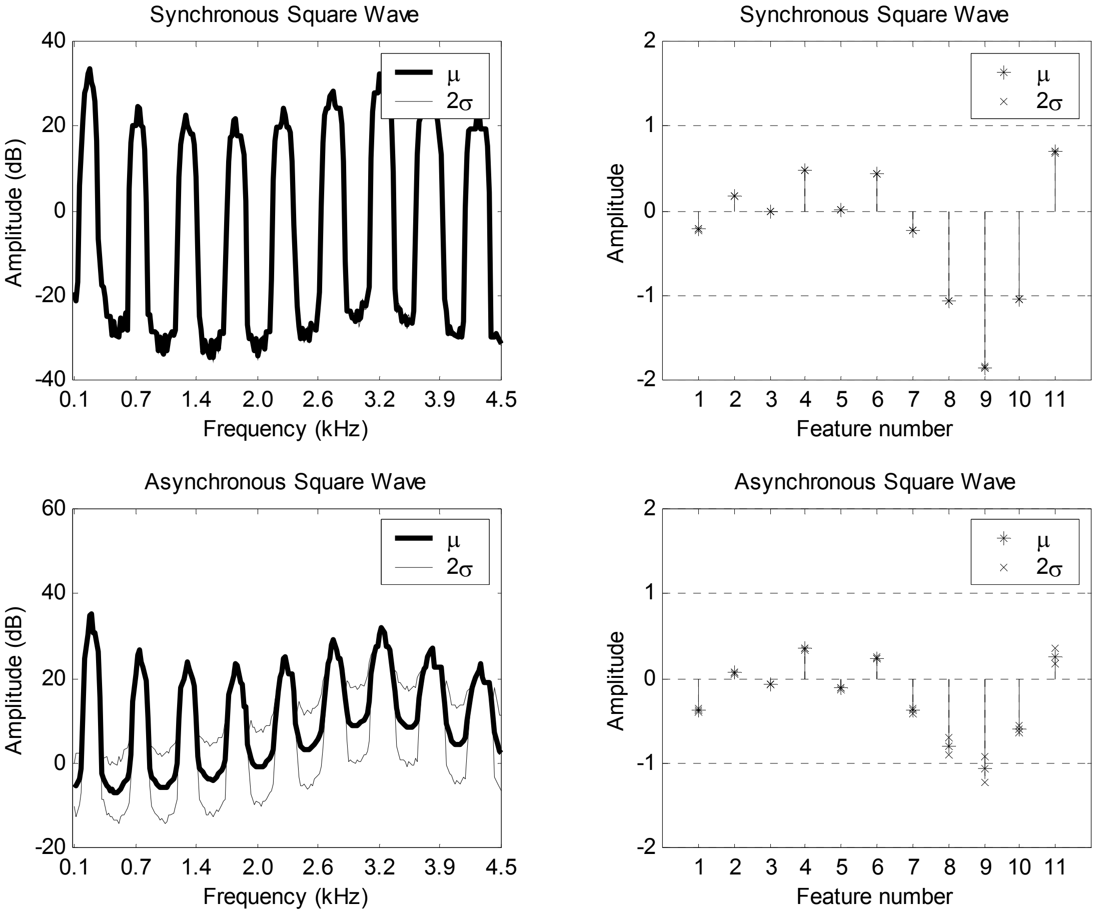


Figure 13: Spectral and feature stability of synchronous and asynchronous square waves, direct signal processing (no sound cards)

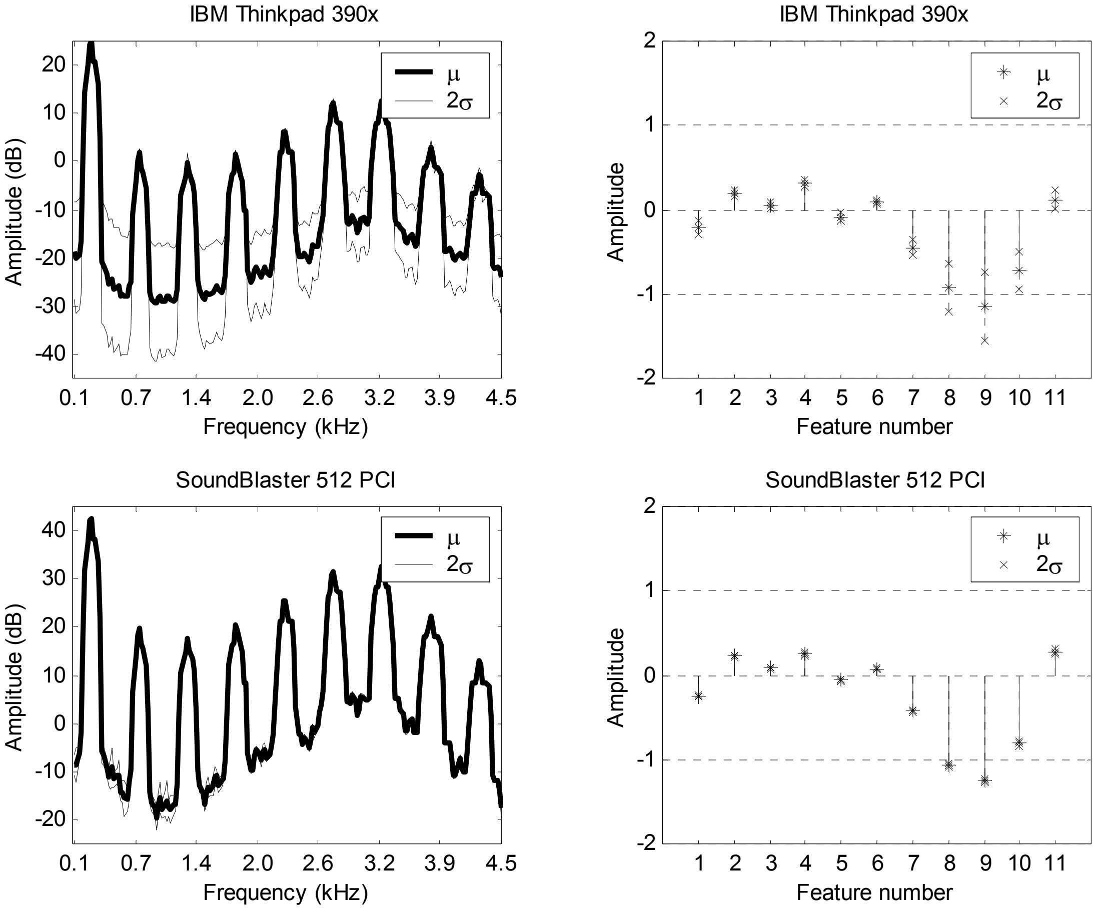


Figure 14: Spectral and feature stability of asynchronous square wave signal using sound cards

DISCUSSION OF RESULTS

The results for the control case (top two panels of Figure 13), with "synchronized" frames indicates very stable spectral and feature plots, as expected. For the case of the control with non-synchronized frames (bottom two panels of Figure 13), there is some variability in the spectrum (approximately a standard deviation of 3dB of spectral variability in spectral valleys between harmonics), and also some noticeable variability in the features, especially the higher-indexed ones. Note, however, that the spectral deviations should be compared to the approximately 40dB difference between the mean of the spectral peaks and the means of adjacent valleys.

The results obtained using the two sound cards (Figure 14) indicate very stable spectral and feature plots for the SoundBlaster 512 PCI, and considerable variability for the IBM ThinkPad. Note that these results were obtained with asynchronous processing. The variability for the IBM ThinkPad is even slightly more than for the case of the asynchronous processing given in Figure 13, whereas the results for the SoundBlaster 512 PCI are nearly the same as those for the synchronous processing depicted in Figure 13. Once again, a big reason for the high spectral variability for the IBM ThinkPad is likely related to the lower signal gain in the absence of AGC.

TESTS WITH VOWEL SOUNDS

These experiments were conducted with the vowels /ah/, /ee/, and /ue/, for one male speaker and one female speaker, for each of the four sound cards mentioned above. As a control, these signals were also inserted in the signal processing path without using a sound card. For each case, spectral means and variances, and feature means and variances were computed for a two second interval beginning at one second into the vowel. In all cases, the "standard" frame spacings and lengths were used (fundamental=250Hz, f_s=11025Hz, segment=100ms, frame=30ms) with no attempt to maintain synchronization of the frames with the signal.

RESULTS

The results for the tests with the vowel sounds are given in figures 15 through 22. The first two figures (15 and 16) depict the spectral and feature stability when the speech signal is processed directly (no sound card), For each of the vowel tokens uttered by the two speakers. The last six figures (17-22) show spectral and feature stability for the two primary sound cards of interest (SoundBlaster PCI512 and IBM ThinkPad 390X).

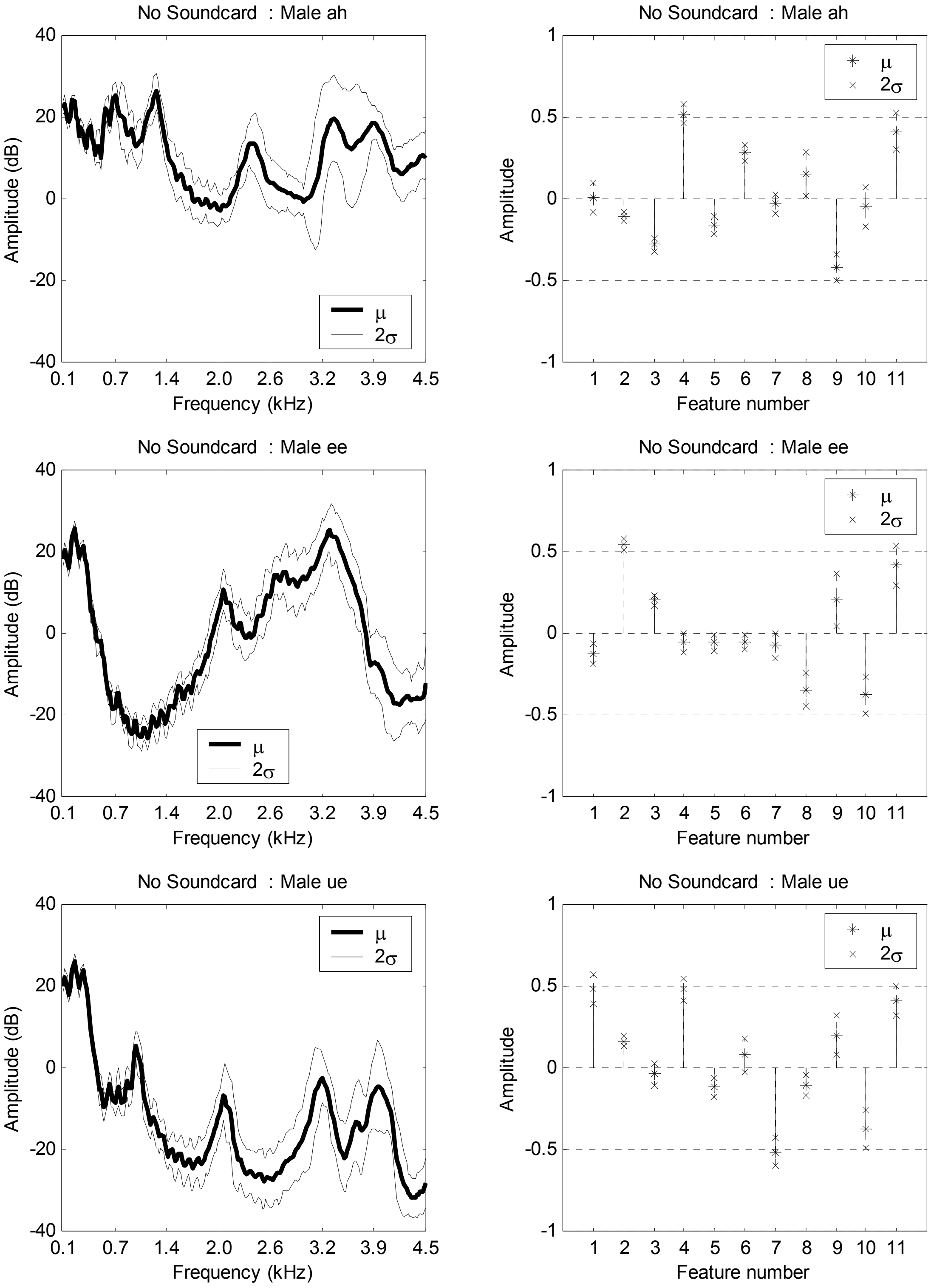


Figure 15: Spectral and feature stability for male speaker /ah/, /ee/ and /ue/ tokens, bypassing soundcard

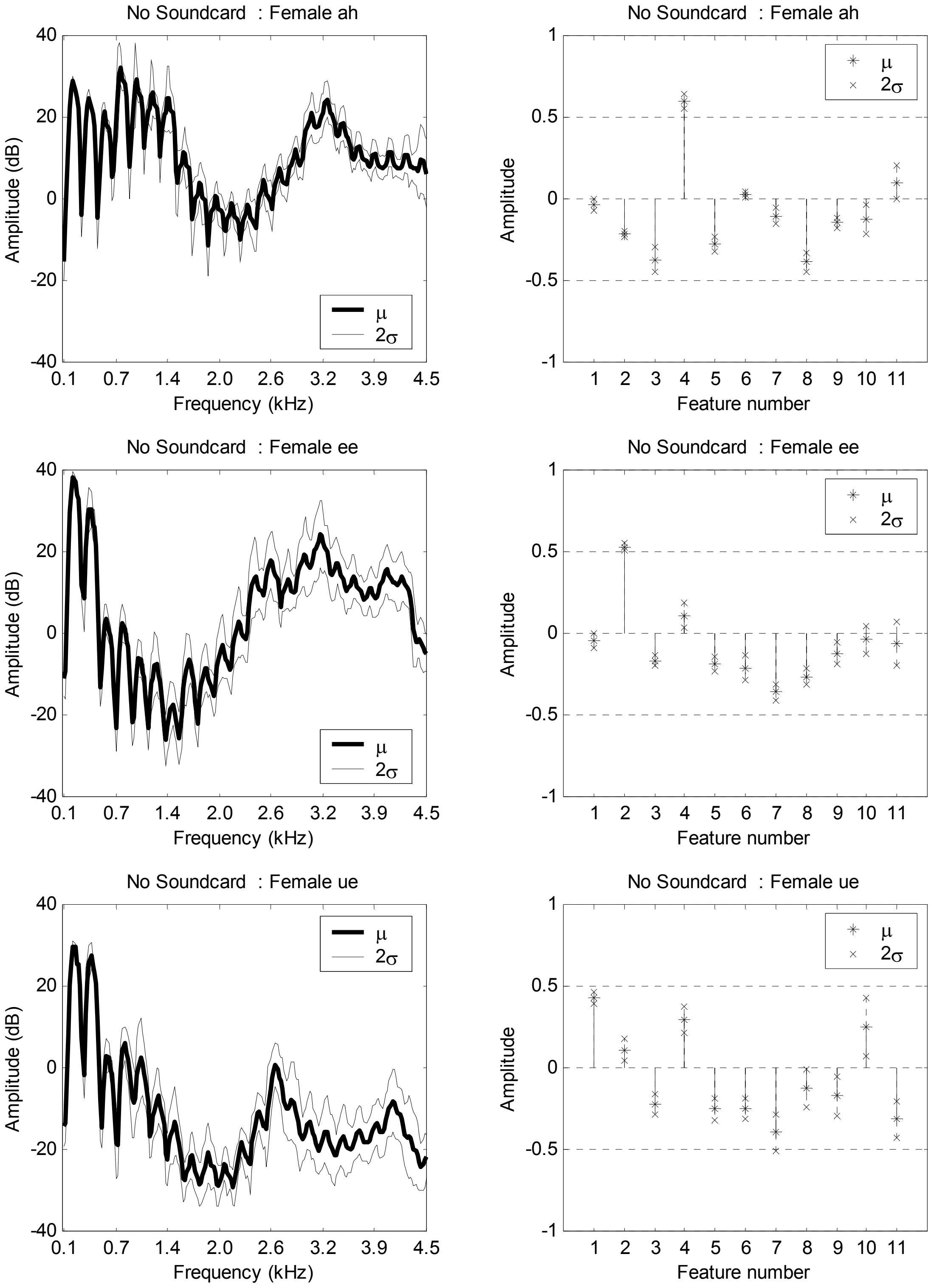


Figure 16: Spectral and feature stability for female speaker /ah/, /ee/ and /ue/ tokens, bypassing soundcard

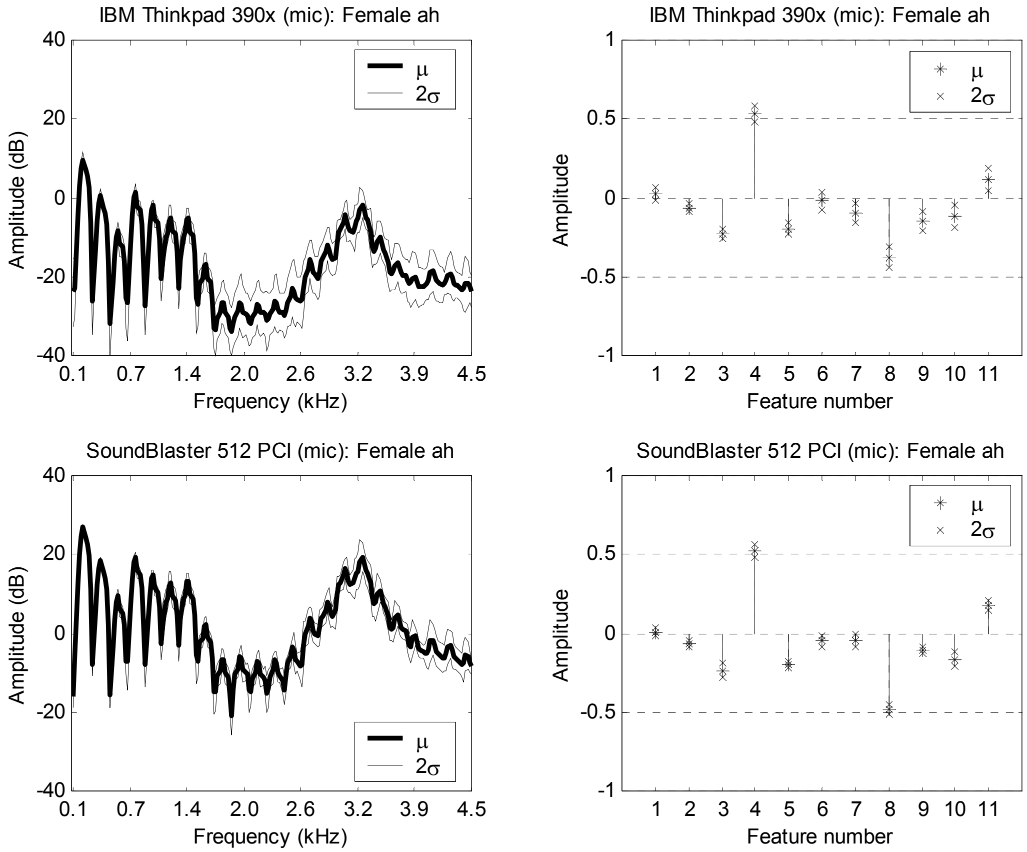


Figure 17: Spectral and feature stability of female speaker /ah/ token using two sound cards

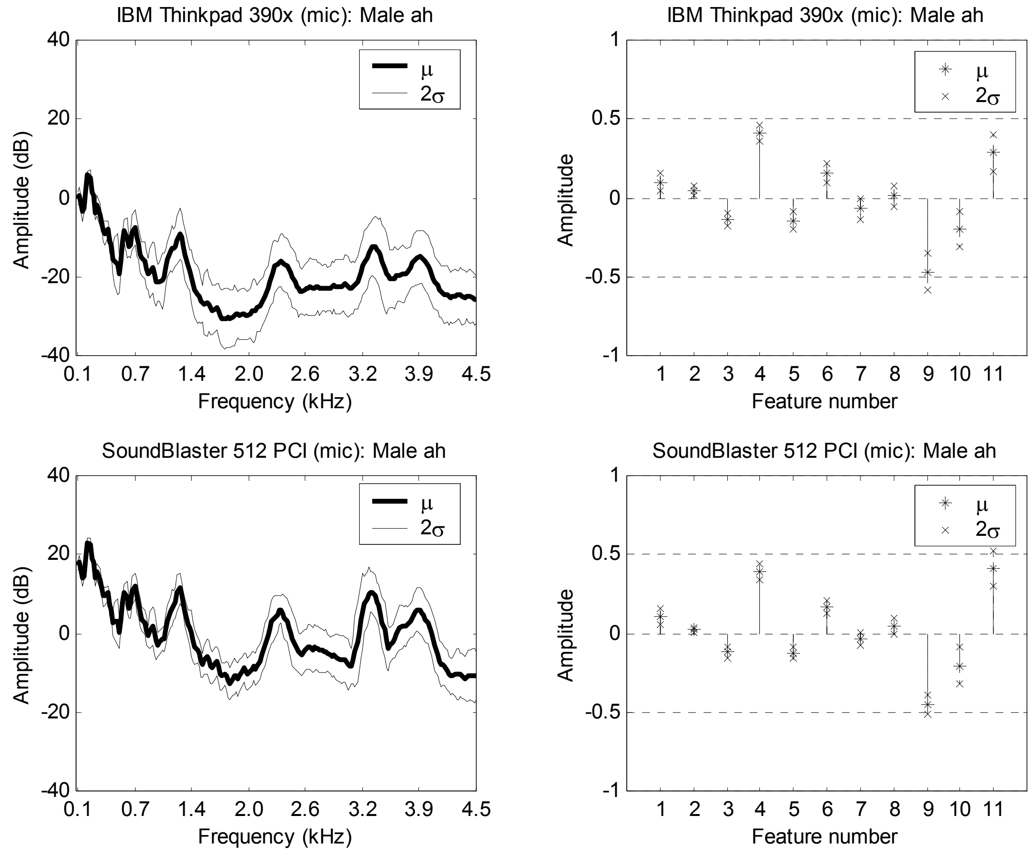


Figure 18: Spectral and feature stability of male speaker /ah/ token using two sound cards

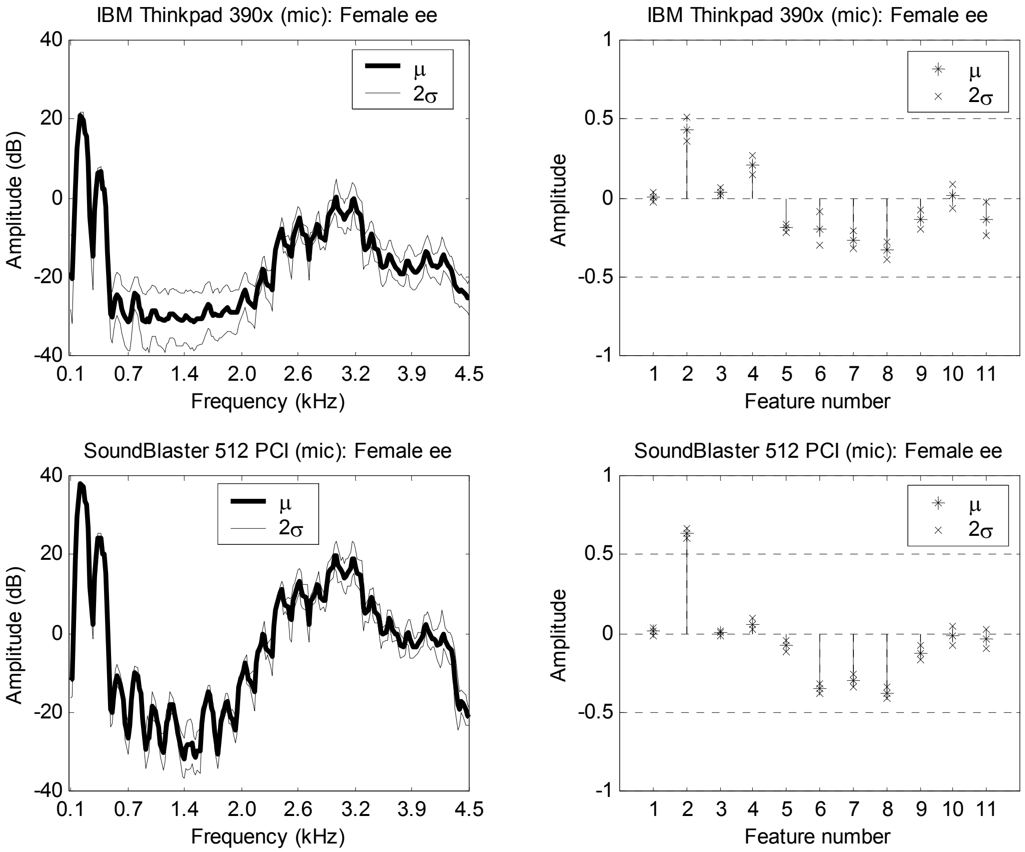
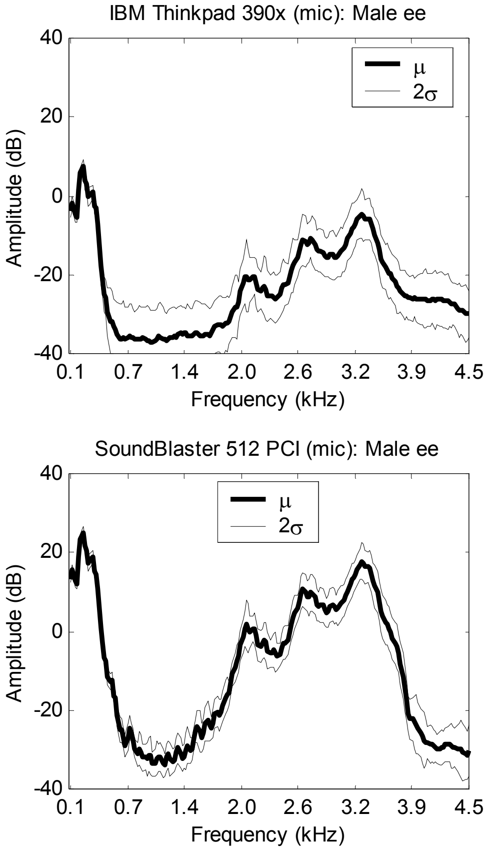


Figure 19: Spectral and feature stability of female speaker /ee/ token using two sound cards



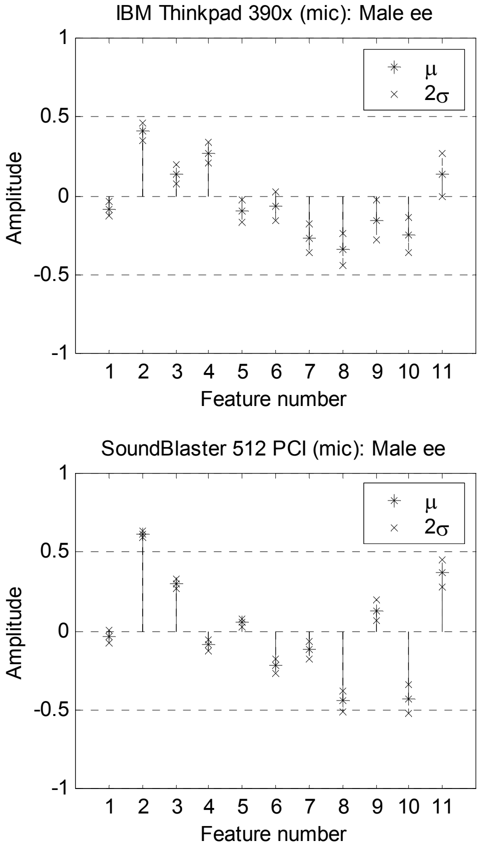
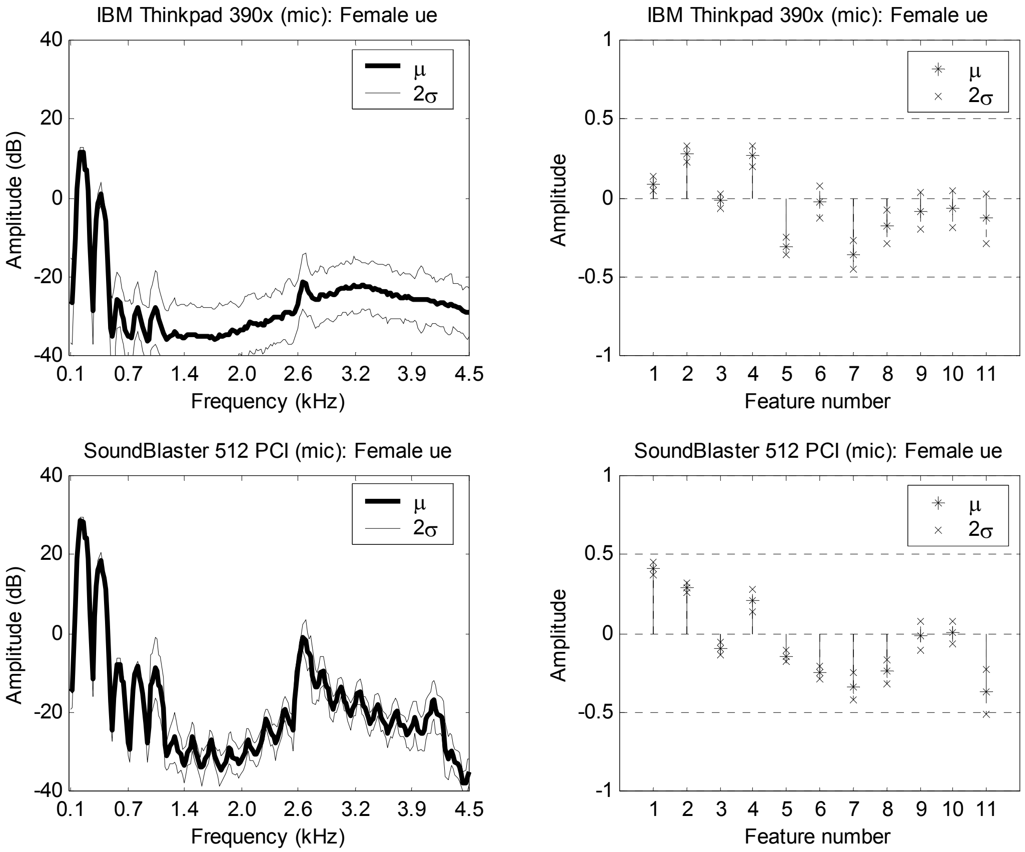


Figure 20: Spectral and feature stability of male speaker /ee/ token using two sound cards



IBM Thinkpad 390x (mic): Female ue

IBM Thinkpad 390x (mic): Female ue

SoundBlaster 512 PCI (mic): Female ue

SoundBlaster 512 PCI (mic): Female ue

Figure 21: Spectral and feature stability of female speaker /ue/ token using two sound cards

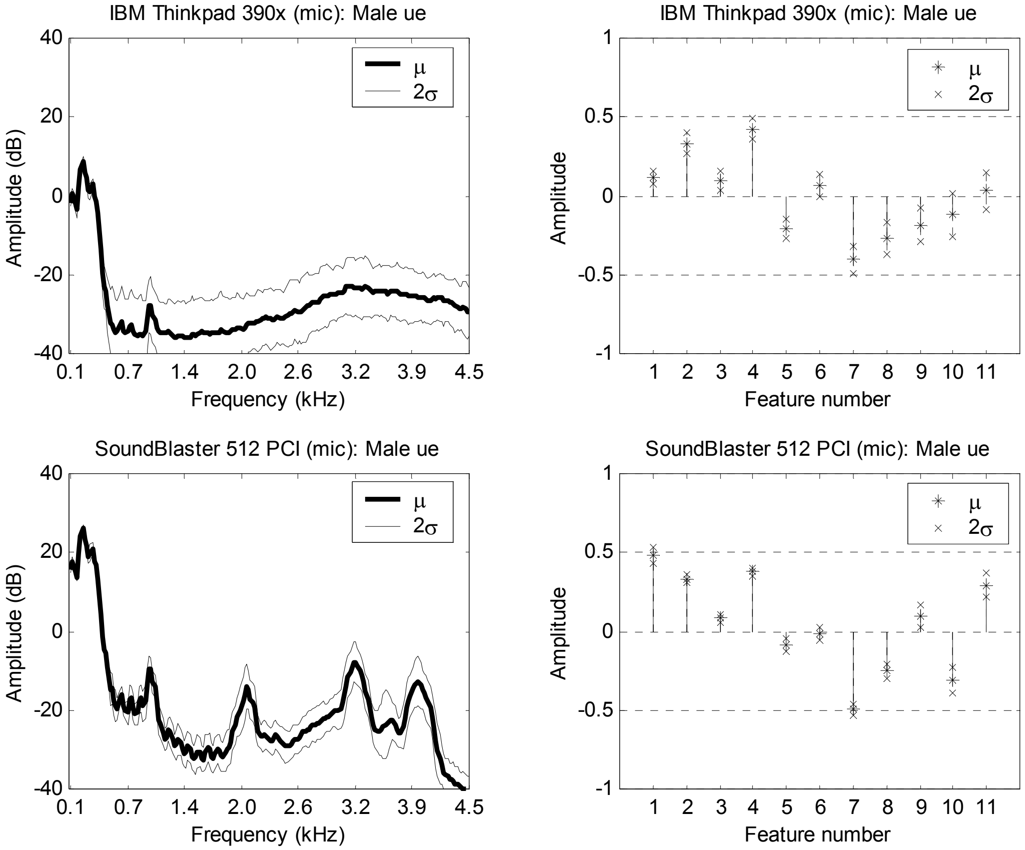


Figure 22: Spectral and feature stability of male speaker /ue/ token using two sound cards

DISCUSSION OF RESULTS

Inspection of the figures reveals that the spectral and feature stability for all vowels in both speakers is virtually the same for the cases where no sound card and a low-noise sound card is used. That is, the PCI512 card appears to introduce no appreciable spectral variability. On the other hand, the IBM ThinkPad 390X did appear to cause additional spectral and feature variability. One of the main differences that can be noted from this sequence of figures is that the female speaker’s spectrum had considerably less jitter than the male speaker’s. This difference appears to be related to the voice quality of the speakers rather than the artifacts introduced by the sound cards.

Summary

The main conclusions from the tests reported in this chapter are as follows:

(1) All sound cards tested appear to have reasonably flat frequency responses across the primary frequencies of interest (about 100 Hz to about .4 Fs) for microphone and aux input. However, the overall frequency response when using the microphone input will of course depend heavily on the response of the microphone.

(2) The lowest quality noise card tested (IBM ThinkPad 390X) did appear to introduce substantially more variability in the signal (more variability in spectral and feature plots) than any of the other cards tested. However, this effect may have been related to the overall lower gain in this card, and lack of automatic gain control.

(3) The two intermediate quality cards (Ensoniq and SoundBlaster 16) appear to have similar overall performance, and were only slightly inferior to the SoundBlaster PCI512.

(4) The variability in the spectrum did not appear to be due to code errors, since both the square wave spectral plots and feature stability plots exhibited low variability when synchronous signals were processed. The variability observed for control signals appeared to be due to window effects introduced when non-synchronous signals are processed.

(5) There were large differences in spectral and feature stability for the two speakers examined.

In conclusion, the three main sources of speech spectral variability for steady state vowel sounds appeared to be (in order of importance): (a), the signal itself, (b) the sound card, (c) and windowing effects. The sound card is only a major factor in speech spectral variability when a very low quality (i.e., high noise) soundcard is used.

CHAPTER IV

MAXIMUM LIKELIHOOD CLASSIFIER

Introduction

During testing of the VATA system, it became apparent that occasionally sounds which were not "correct" examples of any of the 10 steady-state vowels for which the system was designed caused "false" correct displays. For example, ambient room noise could often trigger the display for the /ee/ phoneme. This unintended behavior could cause confusion and frustration for a user relying solely on VATA for pronunciation feedback.

The problem occurs as follows: at the end of each audio segment, the NN issues a decision as to which of the vowel categories the audio stream represents. Unfortunately, the NN must choose from among only the vowel categories for which it was trained; it cannot choose an "other" category. This has the unfortunate consequence of occasionally producing feedback corresponding to a "correct" pronunciation when in fact no valid vowel sound is in the audio stream.

To reduce the number of "false correct" displays, a modified Euclidean-Distance Maximum Likelihood Classifier (MLC) has been placed in tandem with the neural network classifier as a secondary verification system. The MLC essentially measures the how close the features of the sound that the speaker produces are to the features of the ‘typical’ features of the phoneme that the neural network has chosen as the ‘identified’ phoneme. If the features are judged (by an empirically determined criterion) to be “close enough” to the features of the typical pronuciation of the phoneme, the NN’s choice is accepted and displyed to the user; otherwise, the choice is rejected. The following sections descibe the MLC, its use in the VATA system, and results of tests to evaluate the MLC’s effects on system performance.

General Description of the MLC

In its most basic form, the MLC serves as a distance measure for a multivariate Gaussian feature value $\mathbf{f}$ from the mean value $\bar{\mathbf{f}}_i$ for each feature computed from training data. The general formulation requires the covariance matrix $\mathbf{R}_i$, which cannot be calculated accurately without a sufficiently large quantity of training data.

$$D_i(\mathbf{f}) = (\mathbf{f} - \bar{\mathbf{f}}_i)^T \mathbf{R}_i^{-1} (\mathbf{f} - \bar{\mathbf{f}}_i) + \ln\left|\mathbf{R}_i\right|$$

Equation 2: General formulation of Maximum Likelihood Classifier

If $\mathbf{R}_i$ is assumed to be constant, the $\ln\left|\mathbf{R}_i\right|$ term remains constant and can be ignored. If the features are then assumed to be uncorrelated, meaning the that the off-diagoanal terms in the covariance matrix are zero, Equation 2 reduces to the form shown in Equation 3, which is referred to as modified Euclidean distance:

$$D_i(\mathbf{f}) = \sqrt{\sum_{j=1}^{m} \left(\frac{\mathrm{f}_j - \bar{\mathrm{f}}_{ij}}{\sigma_{ij}} \right)^2}$$

Equation 3: Modified Euclidean distance formulation of Maximum Likelihood Classifier

Finally, based on much previous experience with these DCTC features for vowel classification (for example, Zahorian & Nossair, 1999), it was observed that some of the DCTCs are much more useful for discrimination than others. In particular DCTC3 is typically the most important, followed by DCTC4, followed by DCTC2,and then followed by the higher ordered DCTCs in ascending order. Therefore, the distance measure was further modified with a weighting function as follows:

$$D_i(\mathbf{f}) = \sqrt{\sum_{j=1}^{m} w_j \left(\frac{\mathrm{f}_j - \bar{\mathrm{f}}_{ij}}{\sigma_{ij}} \right)^2}$$

Equation 4: Modified Euclidean distance formulation of Maximum Likelihood Classifier, with weight term *w*

These weights, *w*, are important to use if the "information" contained in the underlying features is not proportional to the feature variances. For the case of the DCTC features for vowel recognition, previous work has shown that the DCTCs do not uniformly contribute to vowel recognition. Based on this earlier work, relative weights of (.82, 1.65, 2.47, 2.47, 2.06, 1.65, 1.24, .83, .41, .41, .41, .21, .21, .08, .08) were empirically determined and then used. The actual weights were proportional to those listed, but normalized such that the sum of the weights was 1.0 (Zahorian & Nossair, 1999).

This modified Euclidean distance measure formulation of the MLC is used in the VATA system to provide additional checking to reduce the number of "false correct" displays. It should be noted that although the features used in the VATA system are not completely uncorrelated, they do have low correlation as shown in studies which indicate that principal components of speech spectra are very similar to the cosine basis vector expansion used in VATA (Zahorian & Rothenberg, 1981). Thus the equation used is "reasonably" correct. It also requires much less computation than the full covariance matrix would require. For these same reasons, most Hidden Markov-Model (HMM) speech recognition systems use a diagonal covariance matrix implementation rather than a full covariance matrix implementation (Rabiner, 1993; Lee, 1988).

MLC in the VATA System

To provide verification that the vowel display is producing accurate feedback, the MLC calculates the distance of the average features for the NN's choice from the actual features on which the NN's choice is based. If the feature distance $D_i(\mathbf{f})$ is within the threshold criterion,

$$D_i(\mathbf{f}) < \alpha\sqrt{m}$$

where *m* is the number of features (typically 10 to 15 for VATA), and α is an arbitrary scale factor used for performance tuning, the neural network's decision is accepted, otherwise it is discarded. If α is too small, the MLC will reject many correct tokens; if α is too large, the out-of-category tokens will still not be rejected. For *m* = 11 or *m* = 12, The typical value for α is approximately 1.2, which was experimentally determined, as described below.

To experimentally verify the operation of the MLC, experiments were done as follows: Features were computed using a database of 10 vowel sounds obtained from three speaker groups, as described below, with each speaker producing each vowel sound three times. Vowels were pronounced as “isolated” words, in response to a computer prompt. The central 200 ms interval section of each vowel was used for processing. A two layer feedforward fully interconnected neural network with sigmoidal nonlinearities was trained as a classifier, using nine vowels (those listed above except for /ur/). The means and standard deviations for the features for these nine vowels were also computed, on a vowel by vowel basis, since these are the features needed for the MLC. The neural network/MLC classifier system was then evaluated using test data from 24 different speakers, and using data for 9 vowels, consisting of all the training vowels, except with /ur/ substituted for /oo/. Thus, ideally the classifier should have correctly recognized 8 vowels, but rejected /ur/ as being out of category.

The experiment was repeated separately for male speakers, female speakers, and male/female speakers combined. 50 training speakers were used for the male speaker case, 50 for the female speaker case, and 80 speakers (40 male, 40 female) for the combined case. There were 24 test speakers for the male case, 24 for the female case, and 48 test speakers for the combined case. In each experiment the false acceptance rate, and false rejection rates were obtained as a function of the parameter α above. False acceptances were considered as instances of the /ur/ being accepted as any one of the vowels. False rejections were considered as instances of any vowels classified correctly by the neural network, but incorrectly rejected by the MLC.

Results are shown in Figures 23, 24, and 25, as false acceptance and false rejections as a function of the threshold value α. As expected for low values of α, the false rejection rate is very high. For high values of α, the false acceptance is high. However, for the case of the male and female speakers considered individually, value of α of approximately 1.5 results in very few false rejections, but still rejects most of the out of category tokens.

Initial A/B testing of the VATA display with the MLC enabled and disabled (The MLC can be enabled or disabled via a menu option) showed that when enabled, it successfully reduced the number of "out of category" sounds classified as correct vowel pronunciations. These preliminary A/B test results suggest that inclusion of the MLC the VATA system will produce more usable feedback to users because of the reduced instances of misleading (false positive) displays.

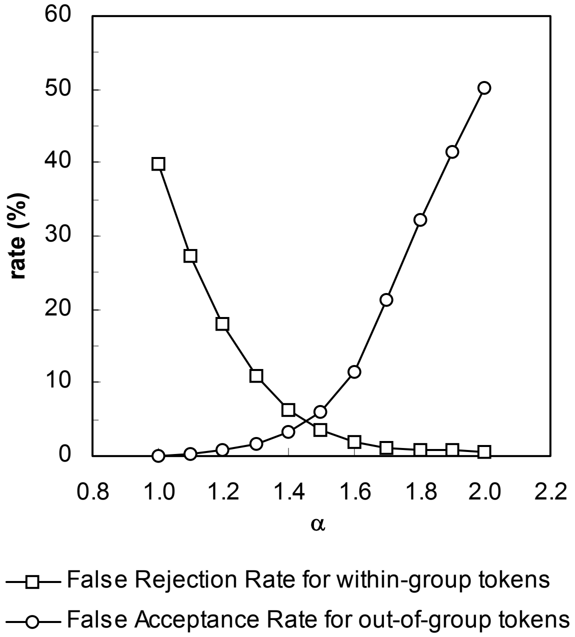


Figure 23: False acceptance/rejection rates as a function of decision threshold value α, male speaker case

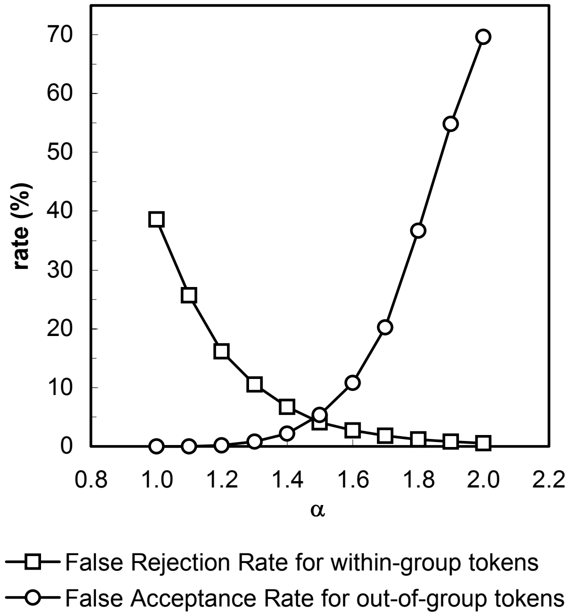


Figure 24: False acceptance/rejection rates as a function of decision threshold value α, female speaker case

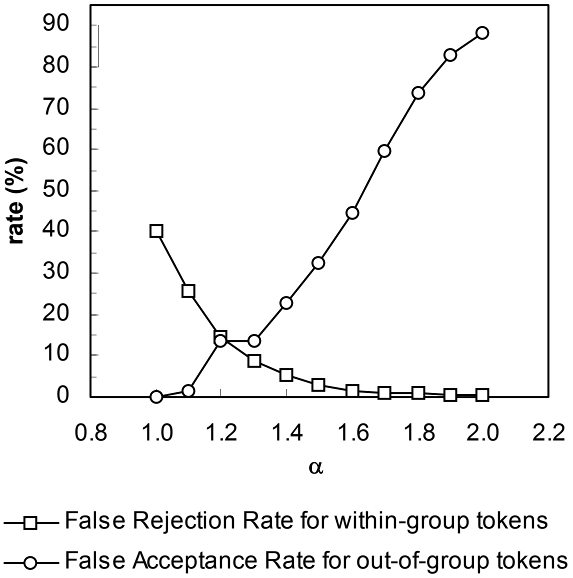


Figure 25: False acceptance/rejection rates as a function of decision threshold value α, combined male/female speaker case

CHAPTER V

PERFORMANCE EVALUATION

Introduction

This chapter presents a detailed performance evaluation of the VATA system. The first section provides results of experiments that show the effects of varying the training database size on the neural network classifier's performance. A second group of tests examines the effects that the use of different sampling rates for the neural network training and test data have on recognition performance. The third group of experiments concern effects high- and low-pass filtering of the speech signal on neural network performance. The last section reports training recognition rate results observed when all available data is used to train the VATA system using the “best” parameters chosen from the previous three sets of results.

Training Database and Hidden Layer Size Variation

As described in the VATA system overview, the neural network classifier used in the VATA system is a fully interconnected feedforward multi-layer perceptron with one hidden layer, trained using the error backpropagation method. Among the many parameters that can affect the performance of this type of network, two of the most significant are the amount of data used to perform the training, and the number of nodes in the hidden layer. Note that the number of inputs must equal the number of features used (typically 11 or 12), and the number of output nodes must equal the number of categories (10, one for each phoneme of interest). Thus the only “free” parameter to determine overall network size is the number of hidden nodes. The amount of training data used can have great impact on the network’s ability to correctly classify an unknown stimulus. The number of hidden layer nodes plays an equally important role in the recognizer’s ability to accurately identify feature patters contained in unknown stimuli as being similar to those contained in the training data.

NEURAL NETWORK TRAINING SET SIZE EFFECTS

To function adequately as a speaker-independent training tool, the VATA system must be trained to recognize correct vowel sounds as produced by a wide variety of voices, and not have its feedback affected by the detailed characteristics of the speaker’s voice or regional accent. By training the neural network recognizer with a large database of speakers which represent a sufficiently wide range of voices and pronunciations of the “correct” phonemes, acoustic and pronunciation differences between voices and accents can be “averaged out,” while the features representing the archetypal utterance can be identified. To this end a large speaker database was assembled as described in Chapter II.

To examine the effects of increasing the number of training speakers in the database, a series of experiments were undertaken in which increasing numbers of speakers were used to train the neural network recognizers. The number of the training speakers varied between one and 100 (males, females, alone) and from two to 200 for males and females combined. For each case a set of 12 test speakers (24 in the combined case of males and females), none of which were included in the training sets, was used to evaluate the performance of the trained network with unknown speakers. Three sets of experiments were run: one for adult males, one for adult female speakers, and one set where the male and female data sets were grouped together.

NEURAL NETWORK HIDEN LAYER SIZE EFFECTS

Another parameter that can impact the VATA system performance is the number of nodes used in the hidden layer of the neural network classifier. The number of hidden nodes has great impact on the neural network’s ability to generate correct responses for a given problem. If the network has too few hidden layer nodes, the classifier lacks sufficient flexibility to be trained for the input data; in effect the system is too simple to adequately model the decision boundaries needed to describe the relationship between the input and output data sets. Conversely, if too many nodes are included in the hidden layer, the recognition performance for test data may be poor because the classifier does not properly generalize.

An analogy can be made to curve fitting: for a given set of domain and range values that correspond to a 3rd order function, attempting to fit a straight line or parabolic curve to the points will yield a poor model. In this case, the order of the function must be increased to fit the curve. Similarly, if a very high order equation is fitted to the data points, the fitted curve may pass through all of the data points, but may not correctly model the general trend and behavior of the curve.

For each of training set sizes mentioned in the preceding section, three hidden layer sizes were tested: 5 nodes, 25 nodes, and 100 nodes. Results of the experiments are shown in Figure 26, Figure 27, and Figure 28.

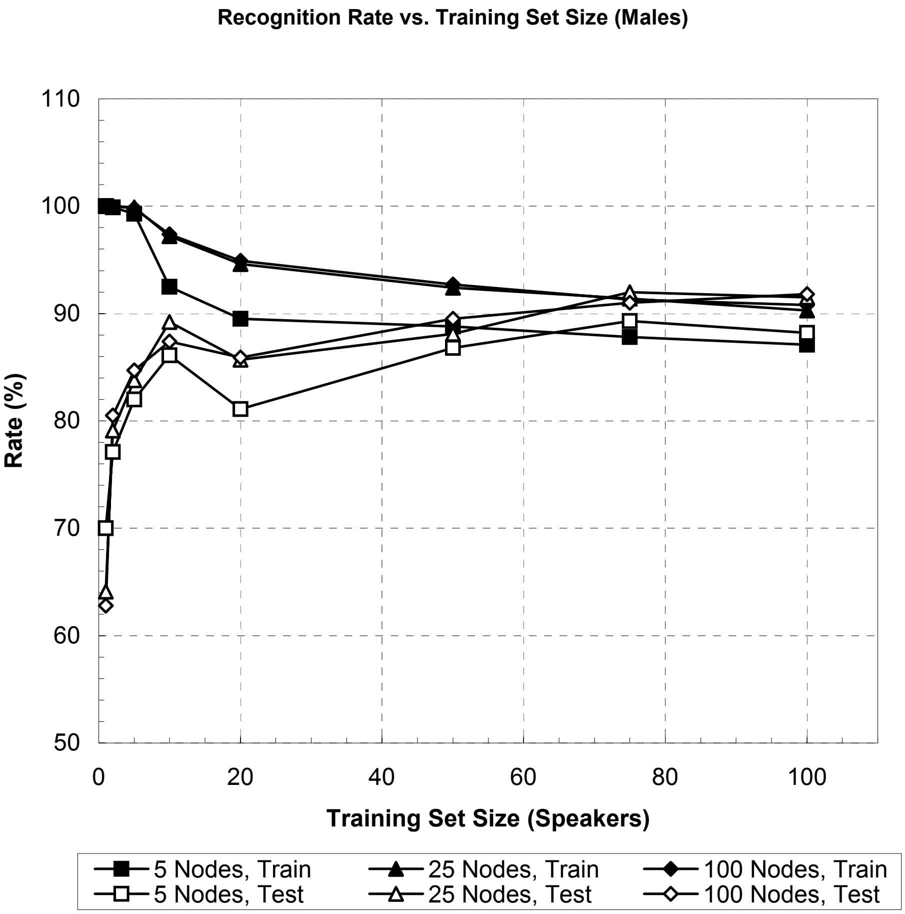


Figure 26: Recognition rate vs. training set size for neural networks with 5, 25 and 100 nodes in hidden the layer (male speakers)

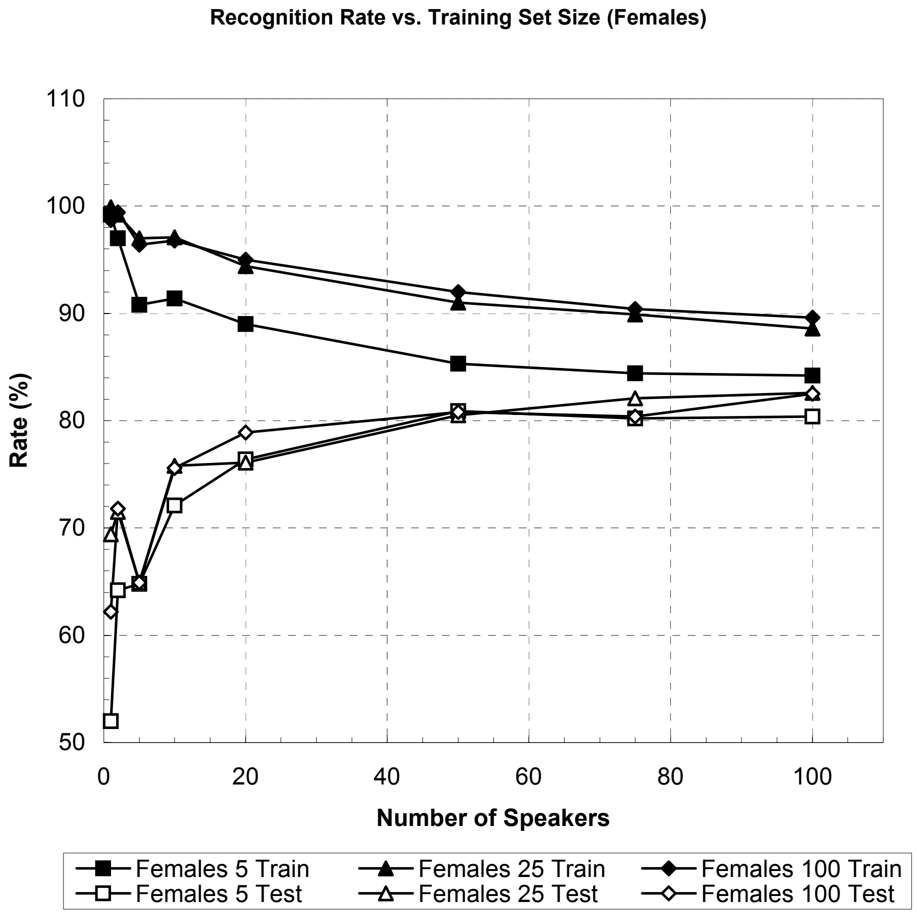


Figure 27: Recognition rate vs. training set size for neural networks with 5, 25 and 100 nodes in hidden the layer (female speakers)

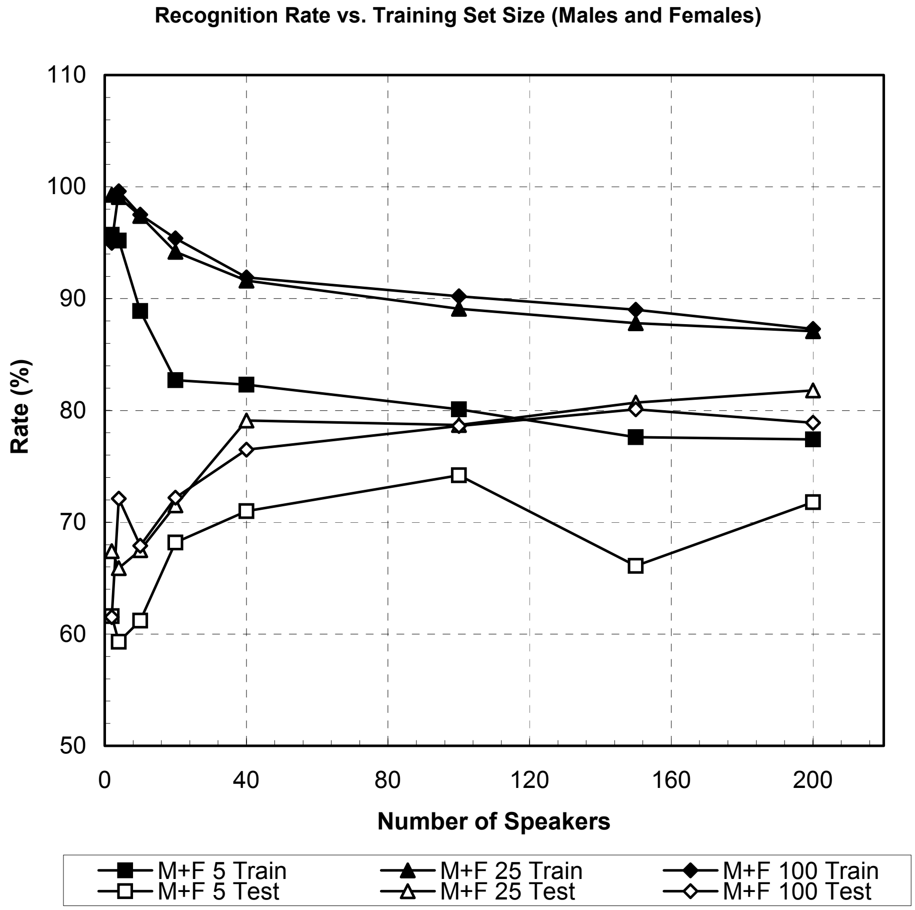


Figure 28: Recognition rate vs. training set size for neural networks with 5, 25 and 100 nodes in hidden the layer (male and female speakers)

DISCUSSION OF RESULTS

Figures 26-28 indicate that, as expected, the training recognition rates tend to decrease as the number of speakers increases and the test recognition rate tends to increase as the number of training speakers increases. More specifically, for the case of the male speakers, the test and training results are approximately equal if 75 or more speakers are used for training. For the female and combined male/female speaker sets, training results consistently remained higher that the test results over the range of speaker set sizes. Of course, these results are influenced by the particular test speakers used. Ideally, even more test speakers would have been desirable or the experiment could have been repeated in “round robin” fashion with rotating sets of test speakers. However, the general trends are quite clear from the results presented.

In terms of the hidden layer size effects, the training results for the larger networks were nearly always higher than the training results for the smaller networks. However the difference between 5 and the 25 node network results was consistently much greater than the difference between the 25 and 100 node networks. This result suggests that a network size of 25 nodes is sufficient for this classification task. For the most part, the networks with 25 and 100 nodes yielded consistently better test recognition rates than the 5 node network, again with relatively small differences between the results of for the 25 and 100 node cases. That is, the larger networks seem to generalize better for the training data to the test data, even with a relatively small amount of training data.

Sampling Rate Effects

Over the many years spanning the course of the VATA system development, advances in computer technology and ever decreasing costs of hard disk storage made possible a transition from a 11 kHz sampling rate to a 22kHz sampling rate. Additionally, the current standard in the automatic speech recognition community is 16kHz minimum sampling rate for speech analysis of “full bandwidth” speech signals. Due to these factors, the standard sampling rate for the VATA database was increased from 11kHz to 22kHz. However, it was desired to make continued use of the speech samples obtained from the 195 speakers recorded at the 11kHz sampling rate. Consequently, a series of experiments were performed to examine the effects of combining speech data of different sampling rates. More specifically, these experiments examined the various combinations of 11kHz and 22kHz speech samples used for testing and training data. All four possible combinations were tried. In each case 28 speakers were used for training and 12 for testing. (For the combined male/female case, the number of both training and test speakers was doubled). Note, however, that the frequency range used for parameter extraction was held at 5kHz in all cases.

Results for all four cases are presented in Table 5. In general, the test results are higher if the training and test sampling rates match. The mixed-case results show a performance drop from between 5 to 15 percent, versus the matched case results. Between the mixed cases, the 22kHz training/11kHz test case yielded slightly better recognition rates than the case using 11kHz for training and 22kHz for testing. This suggests that ideally the VATA system should be used in “real-time” mode with a sampling rate that matches the sampling rate of the data used to train the system. Additionally, the data suggest that if a higher sampling rate is desired for system operation, training data with a sampling rate equal to or greater than the desired rate should be used to train the system. In any case, all new data for the VATA training database is being collected using the 22kHz sampling rate, and all available data (both 11 and 22kHz) is being used to train the system.

Table 5: Training and test recognition rates for different combinations of 11kHz and 22kHz sampled data.

	Train (11kHz)	Test (11kHz)
Male	96.6	86.7
Female	96.7	89.3
Both	93.4	84.0

	Train (11kHz)	Test (22kHz)
Male	96.6	75.4
Female	96.7	65.7
Both	93.4	68.2

	Train (22kHz)	Test (11kHz)
Male	95.6	79.3
Female	93.3	76.6
Both	91.2	79.8

	Train (22kHz)	Test (22kHz)
Male	95.6	89
Female	93.3	82.8
Both	91.2	82.5

Low-pass and High-pass filtering effects

As a further investigation of the changeover in sampling rate from 11kHz to 22kHz, a series of experiments were run to examine the changes in vowel recognition accuracy for both low- and high-pass filtered speech as a function of cutoff frequency in both cases. The experiments were done for both adult male and adult female speakers individually. The number of training speakers used was 28 (Male), 50 (Female) and the number of test speakers was 12. All samples used were sampled at 22kHz, so that frequency ranges up to 10kHz could be tested. The actual filtering was accomplished by selecting the frequency range parameters in the signal processing of VATA, which selects the range of FFT samples used for feature analysis. Results for low-pass filtered speech are shown in Figure 29, and high pass filtered speech in Figure 30.

Recognition performance only appears to degrade only for low-pass cutoff frequencies below 3kHz, and high-pass cutoffs above 100Hz. In other words, virtually all pertinent vowel information appears to be contained in a passband between 100Hz and 3kHz. Thus for the case of vowel recognition, it does not appear that a change to 22kHz sampling rate is necessarily beneficial. Nevertheless, the higher sampling rate could still be beneficial as the system is adapted for use with consonant sounds.

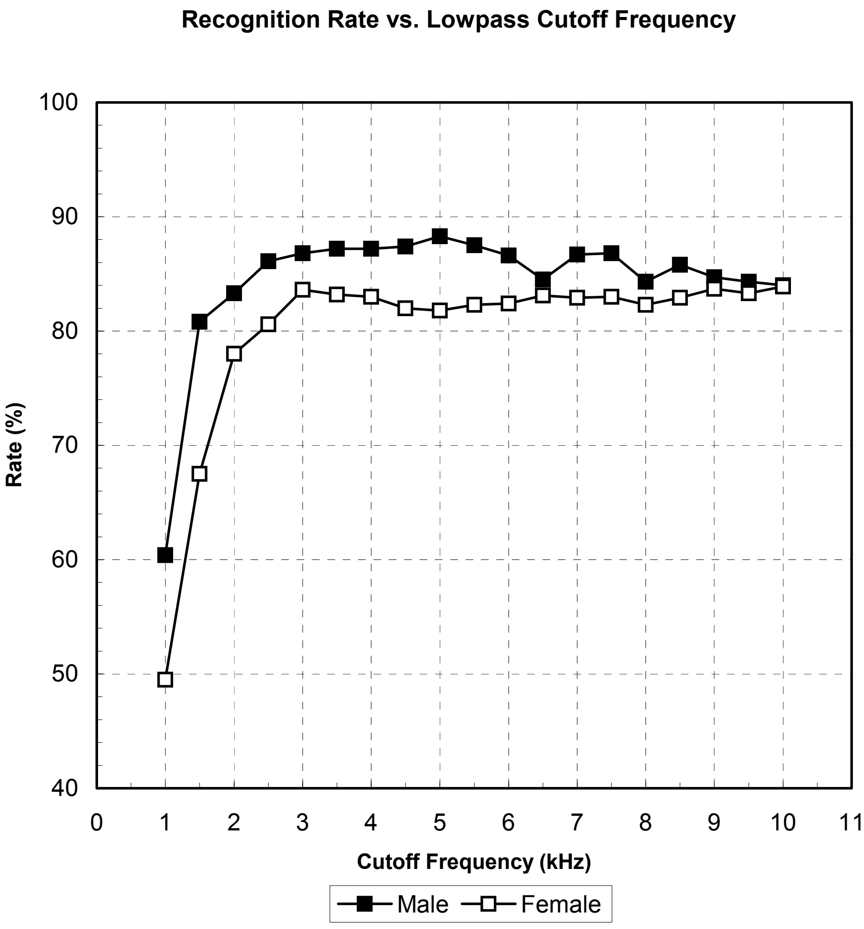


Figure 29: Recognition rate for lowpass filtered speech as the bandwidth increases (lowpass cutoff frequency increases)

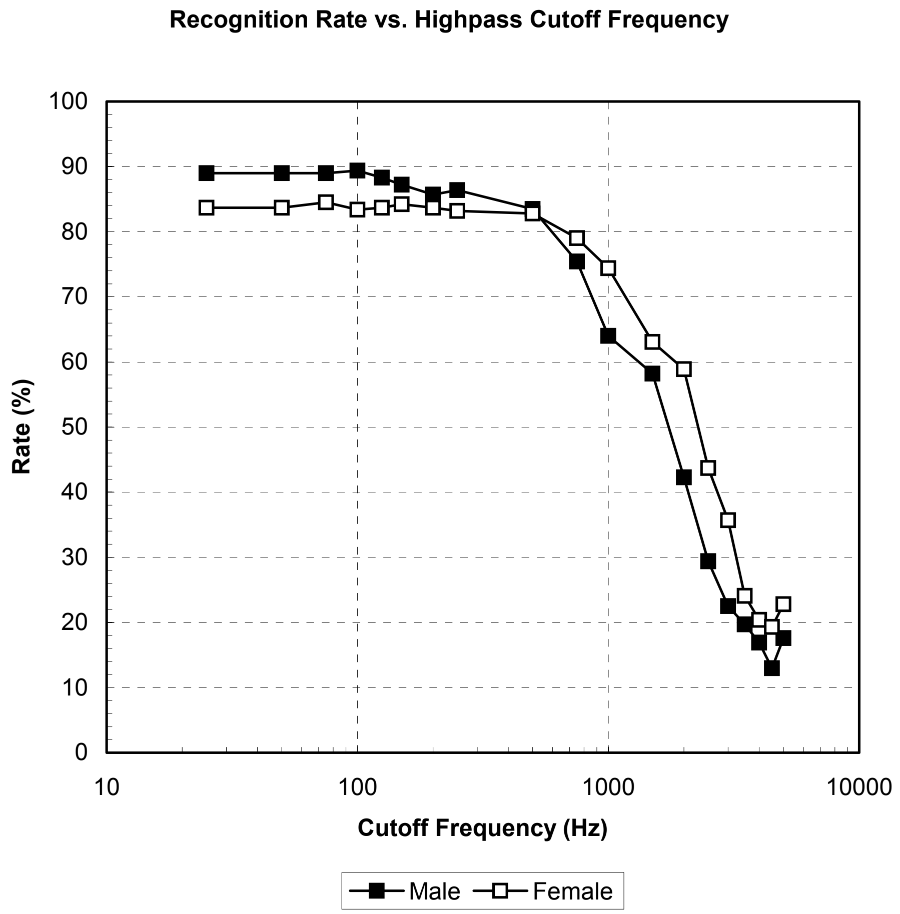


Figure 30: Recognition rate for highpass filtered speech as the bandwidth decreases (highpass cutoff frequency increases)

System Training Results

The fourth set of experiments trained the neural network recognizer using all available data for adult male, adult female, child, and general (data from the three other groups combined) speakers. The results are presented in Table 6.

Table 6: Training recognition rates when all available data is used for training

Speaker Group (speakers in group)	Recognition Rate (%)
Males (114)	89.99
Females (138)	87.33
Children (62)	83.93
General (314)	84.56

Results range from approximately 84 to 90 percent. This set of networks has been incorporated in the operational version of the VATA system. These results constitute the expected upper limits of the system’s accuracy when it is used as a speech training system.

Summary

The results of several experiments investigating various factors which affect the VATA system’s accuracy were reported. These factors include the amount of training data, size of the neural network, sampling rate of speech data, and frequency range effects. A final training experiment was conducted to maximize the expected performance for the current vowel database.

CHAPTER VI

CONCLUSIONS

A new implementation of the Vowel articulation training aid (VATA) has been developed. This implementation contains several improvements over previous versions of the system, including a modular, more extensible Win32-based software architecture, a larger vowel database for system training and testing, and incorporation of a maximum-likelihood classifier to improve system rejection of out-of-category sounds.

In addition to the system improvements, results from several sets of experiments which tested variables which have great impact of the system's performance were reported. These experiments included examination of several points:

- a detailed examination of frequency response and noise properties of sound cards which could potential impact performance for the VATA system
- the effect of training set size on recognition rate
- the hidden-layer size of the neural network classifier
- effects of using different sampling rates for training and test data
- the impact of high-and low-pass filtering of the speech signal on recognition rate
- expected system performance when trained with all available speech data

From the investigation and results of these experiments, reported in the preceding chapters of this paper, several conclusions may be drawn:

- The VATA system appears to work well with all but the lowest quality (highest noise) soundcards, and therefore should be compatible with most consumer-grade sound cards found in the typical multimedia PC system.
- Inclusion of a properly-tuned maximum likelihood classifier aids the VATA system in rejecting out-of category sounds.
- The database used to train the VATA system should be as large as possible; however, the greatest performance gains are seen as the number of speakers approaches 60 for single-speaker type groups. This number is estimated to be higher for multi-speaker type groups.
- A 25-node hidden layer is adequate for the VATA system. Larger networks improve recognition rates slightly, but performance gains are negligible and would not appear to offset negative effects of the increase in computational complexity.
- The best performance is seen when sampling rate of training data matches sampling rate used for real-time operation.
- There is no obvious performance benefit to the VATA system obtained by increasing the training and operation sampling rates from 11kHz to 22kHz.
- The frequency range that contains nearly all vowel information is from 100Hz to 3kHz.
- When all available vowel data is used to train the VATA system, the upper bound on system accuracy is estimated to be between 84 and 90 percent, depending on the speaker group.

REFERENCES

Auberg, S. (1996). Help file for feature computation, unpublished speech lab documentation, Old Dominion University, Norfolk, VA.

Auberg, S. (1996). Speech feature computation for visual speech articulation training. Unpublished Masters Thesis, Old Dominion University, Norfolk, VA.

Beck, A. (1992). Neural network based vowel training aid for the deaf. Unpublished Masters Thesis, Old Dominion University, Norfolk, VA.

Correal, N. (1994). Real-time visual speech articulation training aid. Unpublished Masters Thesis, Old Dominion University, Norfolk, VA.

Dai, B., (1998). Variability analysis of discrete cosine transform coefficient (DCTC) features for speech processing. Unpublished Masters Thesis, Old Dominion University.

Evangelos, S, Fakotakis, N, & Kokkinakis, G., (1991). Fast endpoint detection algorithm for isolated word recognition in office environment, International Conference on Acoustics, Speech, and Signal Processing, pp. 733-736.

Haykin, S., (1993). Neural networks, a comprehensive foundation. New York: Macmillan.

Lee, K. F., (1988). Large vocabulary speaker-independent continuous speech recognition: the SPHINX system. Unpublished Doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA.

Parsons, T. W.,(1987). Voice and speech processing. New York: McGraw-Hill Book Company.

Peterson, G. E., & Barney, H. L. (1952). Control Methods Used in a Study of the Vowels, J. Acoust. Soc. Am. 24, pp. 175-184.

Rabiner, L. & Juang B.,(1993). Fundamentals of speech recognition, New Jersey: Prentice Hall.

Zahorian, S. A., & Jagharghi, A., (1993). Spectral-shape Features Versus Formants as Acoustic Correlates for Vowels, J. Acoust. Soc. Am. Vol.94, No.4, pp. 1966-1982.

Zahorian, S. A., & Nossair, Z. B., (1999). A Partitioned neural network approach for vowel classification using smoothed time/frequency features, IEEE Trans. on Speech and Audio Processing, vol. 7, no. 4, pp. 414-425.

Zahorian, S. A. & Rothenberg, M. (1981). Principal components analysis for low-redundancy encoding of speech spectra, J. Acoust. Soc. Am., 69, 832-845.

Zimmer, A. M., & Zahorian, S. A., (2000). Discriminative and maximum likelihood classifiers for computer-based visual feedback for speech training for the hearing impaired, World Multiconference on Systemics, Cybernetics and Informatics (SCI 2000/ISAS 2000), Vol. VI, Part II, pp. 475-479.

Zimmer, A. M., Zahorian, S. A., & Dai, B., (1998). Personal computer software vowel training aid for the hearing impaired, International Conference on Acoustics, Speech, and Signal Processing, pp. VI-3625-362.

APPENDIX I: DESCRIPTION OF CURRENT VATA SYSTEM

Brief History of VATA

The first implementations of VATA were primarily hardware-based, using analog filter banks for analysis and custom hardware for signal processing and to produce output on a color television monitor. Later versions of the system used custom hardware and dedicated programmable digital processors for signal analysis, coupled with early personal computers to produce graphical displays. Because of the reliance of custom-built and/or specialized hardware, early versions of VATA were expensive and relatively difficult to maintain. As personal computers increased greatly in power and features while declining sharply in price, it became possible to implement VATA completely in software, increasing flexibility and reducing the reliance on expensive and difficult to maintain hardware.

Newest Features

Currently, the VSD is currently implemented as a graphical computer program which runs on the Microsoft® Windows™ 95/98/NTWin32 (95/98/Me/NT/2000) platform. The program requires only a sound card compatible with Windows and a standard microphone. Several improvements have been made over the previous version to enhance functionality and reduce the amount of work necessary to maintain and improve the program.

MODULAR CODE

The previous version of VATA consisted of two separate programs, "WinBall" and "WinBar" (Auberg, 1996). PC Performance limitations present during the development of these programs forced the programs to be written to maximize execution speed. Unfortunately the programming techniques used to maximize the code's speed often led to difficulty in code maintenance and often rendered the code non-reusable.

With the advent of faster computers and inexpensive memory, the necessity for frugality with respect to code size and memory usage has been reduced. Thus, the program has been restructured to offer a greater degree of modularity, which simplifies maintenance and updates. In the latest version, the bargraph and ellipse displays, along with several new diagnostic displays have been combined into one program. The graphics routines, DSP and A/D routines have been placed into separate code modules that can be reused in other projects.

The graphics library in particular has several features that make it useful in projects in addition to VATA:

Extendability: The library employs an "open framework" so that it can be extended to produce an arbitrary number of different types of plots. Currently, the graphics routines can produce histograms and Cartesian plots. Each plot type has several configuration options to suit it to the particular application.

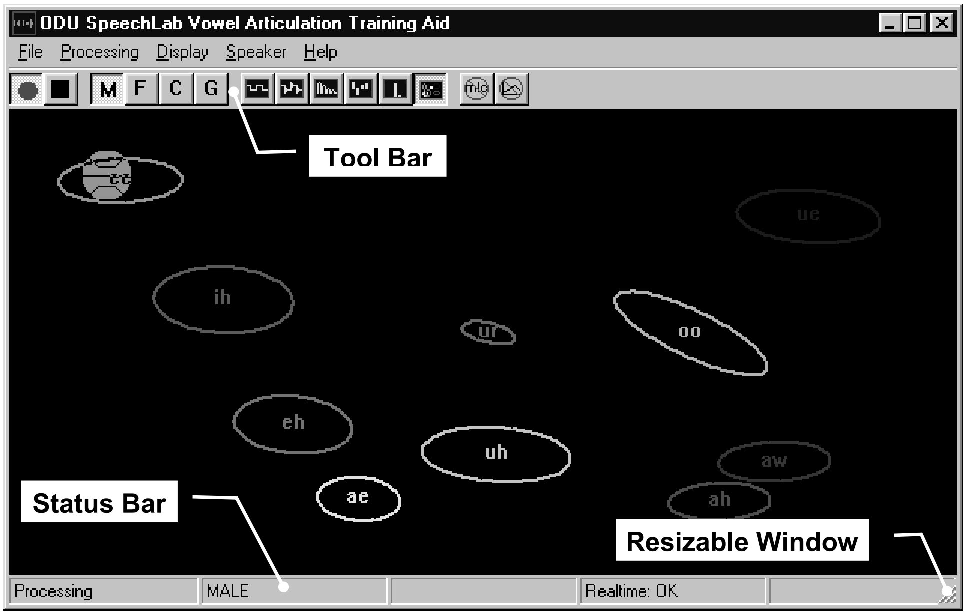
Generic Output: The library creates plots in such a manner that with only minor configuration changes, they may be easily rendered in different colors, in any size, and in several ways, including on screen, by a printer, or in a standard graphics file.

Simultaneous, Multiple Plots: The library can be called any number of times to create multiple plots which can be used simultaneously. Each plot maintains its own internal data and is completely independent from other plots.

Both the A/D code module and the graphics library developed for VATA have found use in other projects, including a graphical user interface for a fetal heart rate monitor system and other internal speech lab applications.

USER INTERFACE IMPROVEMENTS

The Graphical User Interface (GUI) has been completely redesigned from the previous version and now offers standard windows user interface components and behavior.



ODU SpeechLab Vowel Articulation Training Aid

Processing | MALE | Realtime: OK

Tool Bar: A small strip of buttons a the top of the display window, just below the program menu bar give easy access to several program functions and also provide a visual indicator of current settings such as speaker group, display selection, recording activity, and MLC/Prefilter status.

Tool Tips: The tool bar buttons contain small pictograms representing different functions. To assist users in learning the function of each tool bar button, each has a brief description which will appear momentarily when the cursor is placed over the button. These "tool tips" also instruct the user about any shortcut keys (such as 'M' to select the male speaker category) that are associated with the program command.

Status Bar: The status bar at the bottom of the screen shows information pertinent to the usage of VATA, such as current settings and status. Different sections of the status bar show information about Processing status, speaker group selection, and real-time status.

Resizable Window: The VATA display window is fully resizable. This allows VATA to run on a wide variety of PC configurations with different sized monitors and display adapter settings.

On-Line Help: The VATA program now offers an on-line help facility which explains basic usage and documents features accessible via command-line options.

APPENDIX II: DATABASE SUMMARY

The speech database collected during this work contains recordings of steady-state vowels and CVCs produced by more than 300 speakers. The following pages describe the organization and contents of the database from the file level up through the top-level organization of the database CD.

Organization

TOKEN FILES

For each recording ("token") of a specific phoneme, there are two files: the "wave" file, which contains the actual waveform data, and the "phone" file, a simple ASCII-text file which contains label, endpoint and/or segmentation information for the wave file.

The SPHERE file format consists of a 1024-byte-blocked text header, followed by the binary-encoded waveform data. The specific format for the data is encoded in the header. The V/CVC database uses 16-bit Pulse Code Modulation (PCM), with samples stored as signed 16-bit integers contiguously in 'little-endian' format; that is, the low-order byte precedes the high order byte for each sample. A sample header for the SPHERE format used in this database is provided below:

```
NIST_1A
1024
database_id -s3 VOW
database_version -s3 0.1
utterance_id -s18 mATTOOK2_bag____00
channel_count -i 1
sample_count -i 36621
sample_rate -i 22050
sample_min -i -16428
sample_max -i 13108
sample_n_bytes -i 2
sample_byte_format -s2 01
sample_sig_bits -i 16
microphone_id -s4 SLS3
end_head
```

The file header contains the pertinent information for decoding the sample data, some basic information about the token, as well as additional information about the specific token and the database of which it is a part in human-readable text.[*]

The phone file contains information about the phonetic content of the wave file, and how those segments are organized in the token. For example, phone file a typical recording of the CVC word 'bag' looks like this:

```
0 1301 h#
1301 3810 b
3810 6319 ae
6319 8828 g
8830 11181 #h
```

The numbers at the left of each line give the starting and ending sample indices of a phoneme. The letter code following the number is the label for the phoneme found between those sample indices. The labeling system used is based on that used for the DARPA TIMIT database.

Complete specifics of the SPHERE file format are beyond the scope of this paper but can be obtained from the National Institute of Standards and Technology's (NIST) Spoken Natural Language Processing Group.

SPEAKER DIRECTORIES

During a speaker's recording session, the audio files created are stored in a directory for that speaker. The directory is named according to the form:

xIIInLLL

[*] It should be noted that the text header uses standard UNIX-style convention of following each entry with a single newline character ('\n', ASCII code 0x10), and thus may be displayed differently on systems which do not follow this convention.

Where:
x: speaker group code (m, f, c)
III: three letter ID for speaker (usually initials)
n: instance of letter ID (10 maximum)
LLL: three letter ID for recording site

For example, the directory named

mATT0OK2

corresponds to the first male speaker (out of a maximum 10) with the initials "ATT", recorded at site with the letter code "OK2".

GENDER GROUP AND TOP-LEVEL DIRECTORIES

Speaker directories are in turn stored in a speaker group directory (MALE, FEMALE, and CHILD) corresponding to that speaker's gender group. Four top-level directories: V_0, V_CVC_0, V_CVC_1, and V_CVC_2 comprise the top level of the database structure. These top-level groupings are based on recording parameters applied to each group. Table 7 summarizes the contents of each of the four main database sections.

Table 7: Vowel/CVC database summary (as of Ver. 2.1)

Section	Reps		Speakers				f_s	Length
	Vowel	CVC	M	F	C	T	(Hz)	(ms)
V_0	5	0	1	5	13	19	11025	300
V_CVC_0	3	3	18	11	21	50	11025	Variable
V_CVC_1	3	3	55	47	24	126	11025	Variable
V_CVC_2	3	3	40	75	4	119	22050	Variable
Total:			114	138	62	314		

Each of The top-level directories contain recordings which were made using similar criteria. Each time a major revision to the recording criteria was made, a new top level directory was created. The criteria are:

V_0: Each speaker recorded 5 repetitions of each of the following steady-state vowel phonemes:

/aa/, /iy/, /uw/, /ae/, /er/, /ih/, /eh/, /ao/, /ah/, /uh/

Recordings are 300ms segments taken from the speaker's utterance, beginning when the utterance reached a minimum energy criterion within a 50ms window, as implemented in the WinRec automatic database collection software. No other portion of the utterance is preserved. All recordings are at a sampling rate of 11025Hz and use 16-bit A/D conversion.

V_CVC_0: Each speaker recorded 3 repetitions of each of the following:

Steady-state Vowel Phonemes:
/aa/, /iy/, /uw/, /ae/, /er/, /ih/, /eh/, /ao/, /ah/, /uh/
Dipthongs:
/oy/, /ey/, /ow/
CVCs:
"cot", "beet", "boot", "bag", "bird", "pig", "bed", "dog", "cup", "book", "boyd," "cake," and "boat"

Vowel recordings are 300ms segments taken from the middle of the actual utterance. No silence is present at either the start or the end of the token. CVC recordings contain the token plus from 0 to 50ms of silence at one or both ends of the utterance, and are a maximum of 2000ms long. All recordings are at a sampling rate of 11025Hz and use 16-bit A/D conversion.

V_CVC_1: Contains the same tokens as V_CVC_0. Vowel recordings are full recordings of the speaker's utterance, up to 2000ms in length. Up to 50ms of additional sound (typically silence) is preserved at the start and/or the end of the token. CVC recordings contain the token plus from 0 to 50ms of silence at one or both ends of the utterance, and are a maximum of 2000ms long. All recordings are at a sampling rate of 11025Hz and use 16-bit A/D conversion.

V_CVC_2: Contains the same tokens as V_CVC_0 and V_CVC_1. Vowel tokens are full recordings of the speaker's utterance, up to 2000ms in length. Up to 50ms of additional sound (typically silence) is preserved at the start and/or the end of the token. CVC recordings contain the token plus from 0 to 50ms of silence at one or both ends of the utterance, and are a maximum of 2000ms long. All recordings are at a sampling rate of 22025Hz and use 16-bit A/D conversion.

APPENDIX III: VATA FILES SUMMARY

filename.ext	[target]	used by	Description
???????g.dat	[sentence.dat]	TFRONTC	token list (g=m,f,c,a)
???????x.cpf	[cp_fea13.ini]	TFRONTC	feature analysis params
???????x.tsp	[tfront.dat]	TFRONTC	setup params, file names
???????x.tph	[tphone.dat]	TFRONTC	phone codes to process
???????x.sca	[scale.dat]	SCALE	scaling parameters
???????x.fsp	[transfor.dat]	TRANSFOR	setup params, file names
???????x.fof	[transfor.out]	TRANSFOR	output file names
???????x.nd1	[neural.dat]	NEURAL	nn params for BAR disp
???????x.nd2	[neural.dat]	NEURAL	nn params for ELLIPSE disp
???????x.nt2	[neural.tar]	NEURAL	targets for ELLIPSE disp
???????x.fei	[transfor.dat]	TRANSFOR	input file names (ELL)
???????x.feo	[transfor.out]	TRANSFOR	output file names (ELL)
???????x.fed	[neu_trf.dat]	TRANSFOR	ELLIPSE configurations

Ouptut files: (only present if experiment has been run)

filename.ext	[source]	used by	Description
???????g.SCL	[scale.out]	VSD's	Feature scale factors
???????g.NN1	[neural.001]	WinBar	Neural Network Weights
???????g.NN2	[neural.001]	WinBall	Neural Network Weights
???????g.ELL	[ellipse.dat]	VSD's	Ellipse data, token labels

VITA

for
ANDREW MATTHEW ZIMMER

DEGREES:

Bachelor of Science (Electrical Engineering), Old Dominion University, Norfolk, Virginia,
May 1997

Bachelor of Arts (Psychology), University of Virginia, Charlottesville, Virginia,
May 1992

PROFESSIONAL CHRONOLOGY:

Leitch Incorporated,
Chesapeake, Virginia
Software Engineering Group Leader, April 2000 – Present
Software/Firmware Engineer, August 1999 - April 2000

Department of Electrical and Computer Engineering
Old Dominion University, Norfolk, Virginia
Graduate Research Assistant, May 1997 – August 1999
Research Assistant, December 1995 – April 1997

SCIENTIFIC AND PROFESSIONAL SOCIETIES MEMBERSHIP:

Member, IEEE, 1996 – Present
Member, SMPTE, 1999 - Present

SCHOLARLY ACTIVITIES COMPLETED:

Zimmer, A. M., & Zahorian, S. A., (2000). Discriminative and maximum likelihood classifiers for computer-based visual feedback for speech training for the hearing impaired, World Multiconference on Systemics, Cybernetics and Informatics (SCI 2000/ISAS 2000), Vol. VI, Part II, pp. 475-479.

Zimmer A. M., Zahorian, S. A., & Dai, B., (1998). Personal computer software vowel training aid for the hearing impaired, International Conference on Acoustics, Speech, and Signal Processing, pp. VI-3625-362.