CLASSIFICATION OF VISEMES USING VISUAL CUES

by

Nazeeh Shuja Alothmany

BSc, King Abdulaziz University, 1993

MS, University of Michigan, 1998

Submitted to the Graduate Faculty of

Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2009

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Nazeeh Shuja Alothmany

It was defended on

April, 17th, 2009

and approved by

Ching-Chung Li, Professor, Department of Electrical and Computer Engineering

Luis F. Chaparro, Associate Professor, Department of Electrical and Computer Engineering

Amro El-Jaroudi, Associate Professor, Department of Electrical and Computer Engineering

John D. Durrant, PhD, Professor and Vice Chair of Communication Science and Disorders

Susan Shaiman, PhD, Associate Professor, Department of Communication Science and Disorders

Dissertation Director: J. Robert Boston, Professor, Department of Electrical and Computer Engineering

CLASSIFICATION OF VISUAL VISEMES USING VISUAL CUES

Nazeeh Shuja Alothmany, Ph.D

University of Pittsburgh, 2009

Studies have shown that visual features extracted from the lips of a speaker (visemes) can be used to automatically classify the visual representation of phonemes. Different visual features were extracted from the audio-visual recordings of a set of phonemes and used to define Linear Discriminant Analysis (LDA) functions to classify the phonemes.

. Audio-visual recordings from 18 speakers of Native American English for 12 Vowel-Consonant-Vowel (VCV) sounds were obtained using the consonants /b,v,w,ð,d,z/ and the vowels /ɑ,i/. The visual features used in this study were related to the lip height, lip width, motion in upper lips and the rate at which lips move while producing the VCV sequences. Features extracted from half of the speakers were used to design the classifier and features extracted from the other half were used in testing the classifiers.

When each VCV sound was treated as an independent class, resulting in 12 classes, the percentage of correct recognition was 55.3% in the training set and 43.1% in the testing set. This percentage increased as classes were merged based on the level of confusion appearing between them in the results. When the same consonants with different vowels were treated as one class, resulting in 6 classes, the percentage of correct classification was 65.2% in the training set and 61.6% in the testing set. This is consistent with psycho-visual experiments in which subjects were unable to distinguish between visemes associated with VCV words with the same consonant but different vowels. When the VCV sounds were grouped into 3 classes, the percentage of correct classification in the training set was 84.4% and 81.1% in the testing set.

In the second part of the study, linear discriminant functions were developed for every speaker resulting in 18 different sets of LDA functions. For every speaker, five VCV utterances were used to design the LDA functions, and 3 different VCV utterances were used to test these functions. For the training data, the range of correct classification for the 18 speakers was 90-100% with an average of 96.2%. For the testing data, the range of correct classification was 50-86% with an average of 68%.

A step-wise linear discriminant analysis evaluated the contribution of different features towards the dissemination problem. The analysis indicated that classifiers using only the top 7 features in the analysis had a performance drop of 2-5%. The top 7 features were related to the shape of the mouth and the rate of motion of lips when the consonant in the VCV sequence was being produced. Results of this work showed that visual features extracted from the lips can separate the visual representation of phonemes into different classes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I write these words wishing that my father had lived to see this day. He was my main motivation for pursuing graduate work and he always wished that all of his children would manage to obtain high degrees. His dream became a reality and I hope that his sole will rest in peace.

I would like to thank my advisor Dr. Robert Boston for the support and guidance he provided me throughout my Ph.D. I would also like to express my gratitude to all my research committee members for the time and valuable feedback they have given me.

I would like to thank the Electrical and Computer Engineering Department at King Abdulaziz University in Jeddah Saudi Arabia for providing me with a scholarship to pursue my graduate degree and I look forward for joining the department as a faculty member.

I would like to thank my older brother Dr Dheya Alothmany for his constant support and encouragement for me during my stay in USA. I would also like to thank my mother Sabiha who kept praying for my success day and night.

I would not have been able to complete my work if it had not been for God's blessings first, and then the enormous support of my wife Souad Rahmatullah. She stood by me through difficult times and encouraged me to keep my spirit high. My lovely children Danyah, Hamzah, Mouaz and Zayd can now get ready to go back home. I dedicate this dissertation to all those who contributed towards the success of this work.

1.0 INTRODUCTION

A human being has five senses - sight, hearing, touch, smell, and taste. One can acquire more exact information on the surrounding environment by integrating cues obtained from different senses. This merging of cues from different senses in humans has led many researchers to investigate the effects of combining several modalities or sensor outputs together on the performance of many automated systems currently utilizing a single modality.

Hearing impairment is one of the handicaps common among people, and amplification devices to compensate for it date back several centuries. These amplification devices are often called hearing aids, and their aim is to maximize speech understanding for individuals with hearing impairment. However, some researchers believe that the performance of these devices in noisy environments has not yet reached a satisfactory level that justifies the cost to the patient [1]

Alternatively, a person skilled in lip reading is able to infer the meaning of spoken sentences by looking at the configuration and the motion of the visible articulators of the speaker, such as the tongue, lips, teeth, and cues from the context. Lip reading is widely used by hearing impaired persons for speech understanding. In addition to lip reading, facial expressions and body language can be used to assist in aural communication [2]. Sumby and Pollack [3] showed that adding visual information to acoustic waveforms is equivalent to a 12dB increase in the signal to noise ratio (SNR).

One of the important studies that attempted to investigate the relation between acoustic and visual information was conducted by McGurk [4]. The study showed that presenting a viewer with conflicting audio-visual recordings of a certain word results in the wrong perception of the sound. This study demonstrates that vision can play a role in speech perception. Clavert [5] confirmed the Mcgurk effect by using functional Magnetic Resonance Imaging (FMRI) to show that the speech perception center in the brain analyzes speech-like lip motions even when no sound is present.

It is not yet clear how the brain combines visual information with audio information to understand speech. It is also not clear what kind of visual cues are utilized by the brain in this process. Therefore, many studies have attempted to explore different methods of combining visual cues with acoustic information in addition to utilizing different types of visual cues to represent the speech [6-10].

In some applications, such as lip reading and screening of security camera recordings, there is no access to audio. In other applications, an audio signal might be present but severely corrupted. Having automatic viseme classifiers based on visual cues will help in narrowing down the list of possible spoken phonemes. Furthermore, the identification of visemes might be useful to adjust the parameters of hearing aid filters for better performance in situations where more than one person is talking.

The initial objective of the present study was to investigate whether utilizing visual information in conjunction with a hearing aid device would enhance the performance of the device in noisy environments. To further investigate this objective, audio-visual recordings were obtained from the Advanced Multimedia Lab at Carnegie Mellon University [11]. The data consisted of recordings for the change in x-y coordinates of lip corners together with lower and

upper lip heights in consecutive frames of video recordings for subjects repeating different words. The points of interests on the lips in that database are shown in Figure 1-1. The consistency of the visual information accompanying the utterance of specific words across different speakers was evaluated.

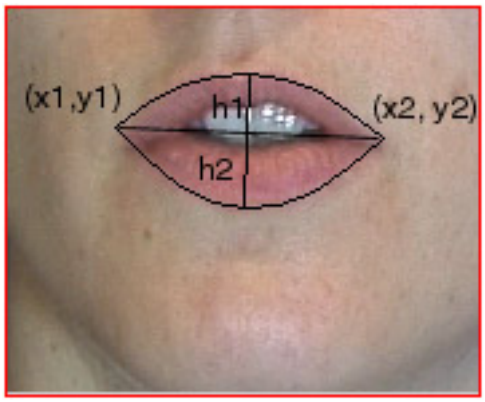

Figure 1-1 Points of focus around the lips in Chen's audio-visual data base

The data set showed that a repeated pattern for the same word exists within the same speaker and even across speakers. However, since these data were for connected speech, it was difficult to quantify and model these patterns. In order to build and systematically evaluate models of lip motion patterns, models should first be built for smaller blocks of speech, then extended to connected speech.

Phonemes are the basic units of speech in the acoustic/auditory domain. The term "viseme" has been introduced as an analogy to represent the visual representation of the phoneme [12]. A viseme describes the particular facial and oral movements that occur with the

voicing of phonemes, and they are considered the smallest distinguishable visual unit of speech [13-17].

Researchers concerned with speech production and lip reading have obtained different viseme-to-phoneme mappings and identified phonemes that are visually confused with one another. One of the important studies conducted in this area was done by Owen and Blazek [18]. The group presented video recordings of vowel-consonant-vowel (VCV) sounds for 23 English consonants to 10 subjects. Subjects observed video sequences of speakers (without audio) saying a certain sound. Based on the motion patterns of the visual articulators (lips, tongue, throat, teeth, and facial emotions) they observed, the subjects attempted to identify the produced sound. Results from all subjects were gathered in matrices with row representing the actual VCV sequence and columns representing the response of the subject. These matrices were called confusion matrices. When a cluster of confused sequences appeared in the matrix, a 75% response criterion was used as a requirement for considering these phonemes to have the same visual representation (i.e. viseme). The viseme-to-phoneme mapping obtained by Owen is consistent with mappings obtained by many other researchers working in speech production as well as audio-visual signal processing as discussed in Section 2.2.

The objective of the current study is to analyze the feasibility of using a set of visual features extracted from the 2D images of lip motion in an automated classification system to classify sounds into different visemes. Different visual features from the audio-visual recordings of a set of phonemes were extracted and used in a stepwise linear discriminant analysis to identify which visual features are most effective in distinguishing between the different visemes.

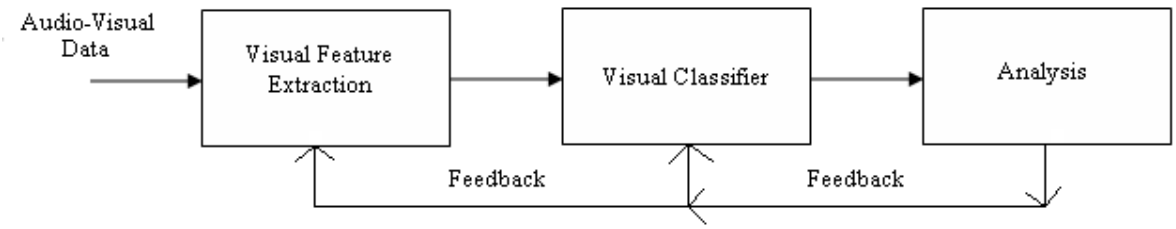

Figure 1-2 Steps of designing an automatic viseme classifier based on visual cues

Figure 1-2 shows a block diagram representing the steps involved in reaching the objective. The figure has three major blocks, with feedback loops between them. The first step in designing a visual classifier of phonemes is to develop an audio-visual database consisting of recordings for speakers uttering different phonemes. Visual features are extracted from the video images and used to train and test a visual classifier. The parameters of the classifier were modified based on the feedback coming from the results of analyzing the classification output.

Chapter Two of this thesis reviews the literature available on different audio-visual applications and then concludes with a summary of the objective of the study that is based on the review. Chapter Three describes the experimental setup. Chapter Four presents the results of the experiments conducted. Chapter Five discusses the results. The conclusion is presented in Chapter Six and suggestions to future work are presented in Chapter Seven.

2.0 LITERATURE REVIEW

There have been a wide range of studies in the area of audio-visual signal processing. In each study, different visual features as well as different classifiers have been chosen and tested for performance. Despite this extensive research, there are many research questions still open in this area, for example [13]

1. Which facial features are important? How are geometric features such as lip height and width related to non-geometric features such as discrete cosine transform of the mouth image?
2. What methods offer an effective means of using visual information and audio information together for speech comprehension?
3. How can visual cues such as face pose and gaze be effectively used to direct the attention of audio speech recognition to enhance the robustness of the audio signal?

This chapter summarizes the different approaches used by many researchers working in audio-visual applications. Section 2.1 discusses work done in studying the correlation between the audio and visual signals. Section 2.2 focuses on work in the area of mapping visemes to phonemes. Section 2.3 presents some miscellaneous audio-visual applications and Section 2-4 summarizes the review and re-states the objectives of this study based on the literature review.

2.1 AUDIO-VISUAL CORRELATION

Studies have been conducted by many researchers to investigate the correlation between visual articulators and sounds produced by them. Many parameters have been used in these studies to represent both visual and audio signals.

Hani *et al* [19] examined the linear association between the vocal tract configuration, facial behavior and speech acoustics. They applied linear estimation techniques to support the claims that facial motion during speech is largely a byproduct of producing the speech acoustics. The experimental data they used included measurements of speech acoustics, the motion of markers placed on the face and in the vocal-tract for two subjects. The numerical results showed that, for both subjects, 91% of the total variance observed in the facial motion data could be explained by vocal-tract motion by means of simple linear estimators. For the inverse path, i.e. recovery of vocal-tract motion from facial motion, their results indicated that about 80% of the variance observed in the vocal-tract can be estimated from the face. Regarding speech acoustics, they observed that, in spite of the nonlinear relation between vocal-tract geometry and acoustics, linear estimators are sufficient to explain between 72 and 85% (depending on subject and utterance) of the variance observed in the RMS amplitude of the spectral envelope.

J Barker [20] showed that there is correlation between the linear estimate of acoustics from lip and jaw configuration and speech acoustics itself. In his study, the lips and jaw movements were characterized by measurements taken from video images of the speaker's face, and the acoustics were characterized using spectral pair parameters and a measure of RMS energy. The speech acoustics estimated from the lip and jaw configurations had a correlation of 0.75 with the actual speech acoustics.

Ezzat and Poggio [21] designed a system that had a set of images spanning a wide range of mouth shapes. They attempted to correlate those images with the phonemes from a speech signal. The purpose was to use the correlation results in animating a lip that moves according to the speech signal. Their system takes input from a keyboard and produces an audio-visual movie of a face enunciating that sentence. The generic name of their system is Mike Talk, and videos of their results are available on their website [22]. Their results indicated a correlation between the lips and the acoustics produced.

Roland *et al* [23] investigated the statistical relationship between the acoustic and visual speech features for vowels. Their study used an audio-visual speech data corpus recorded using Australian English. The acoustic features were the voice source excitation frequency $f_{0,}$ the formant frequencies f_1-f_3, and the RMS energy, while the visual features were extracted from the 3D positions of the two lip corners and the mid point of upper and lower lips as shown in Table 2-1. Several strong correlations are reported between acoustic and visual features. In particular, F1 and F2 and mouth height were strongly correlated.

Table 2-1 Acoustic and visual features used by Roland

Acoustic feature	Visual feature
Voice source excitation f_0	Mouth height
Formant frequency F_1	Mouth width
Formant frequency F_2	Lip protrusion
Formant frequency F_3	

The studies presented in this section indicate that the audio and visual signals are correlated, which justifies the attempt to model the visual representation of phonemes. Most of the studies in this area focused on the visual cues related to the mouth and lips, which led us to focus on visual features related to the mouth area in designing the first block shown in Figure 1-2.

2.2 LIP READING

A person skilled in lip reading is usually able to infer the meaning of spoken sentences by looking at the configuration and the motion of visible articulators of the speaker such as the tongue, lips, and teeth. This skill of lip reading is widely used by hearing impaired persons for speech understanding. However, lip reading is effective only if the speaker is observed from the frontal view. In addition, lip reading becomes difficult if more than one person is talking at the same time because a lip reader can focus only on one speaker at a time.

This section reviews some of the concepts involved in lip reading as well as research that has been conducted in the visual identification of phonemes.

2.2.1 Visemes and Phonemes

The Webster English dictionary defines phonemes as abstract units of the phonetic system of a language that correspond to a set of similar speech sounds which are perceived to be a single distinctive sound in the language [24]. An example of a phoneme is the /t/ sound in the words “tip”, “stand”, “water”, and “cat”. Since the number of consonants in the world's

languages is larger than the number of consonant letters in any one alphabet, linguists have devised systems such as the International Phonetic Alphabet (IPA) to assign a unique symbol to each consonant [25]. The Longman Pronunciation Dictionary, by John C. Wells [24], for example, used symbols of the International Phonetic Alphabet and noted that American English has 25 consonants and 19 vowels, with one additional consonant and three additional vowels for foreign words.

The term "viseme" combines the words "visual" and "phoneme" [12]. Visemes refer to the visual representations of lip movements corresponding to speech segments (phonemes), and they are considered the smallest unit of speech that can be visually distinguished [13], [15], [14], [16], [17].

The mapping between visemes and phonemes is many to many, meaning that one viseme may correspond to more than one phoneme and the same phoneme can correspond to multiple visemes. This happens because the neighboring phonemic context in which a sound is uttered influences the lip shape for that sound. For example, the viseme associated with \t\ differs depending on whether the speaker is uttering the word "two" or the word "tea". In the former case, the \t\ viseme assumes a rounded shape in anticipation of the upcoming \uu\ sound, while in the latter it assumes a more spread shape in anticipation of the upcoming \ii\ sound[18, 26]. Researchers have developed many mappings between the visemes and phonemes, which are discussed in the remaining part of this section.

Faruqui *et al* [27] used a map between Hindi phonemes and 12 visemes, where several phonemes were mapped to one viseme. The mapping shown in Table 2-2 was used to animate a face with lips moving in synchrony with an incoming audio stream. In this system, once the incoming audio signal was recognized, the mapping shown in Table 2-2 was used to select the

corresponding viseme to be animated to the observer. The paper did not explain how this mapping was obtained, but it stated that the animated faces were shown to different observers and the perception rates were promising.

Verma *et al* [28] also used this mapping with a modified scheme for synchronizing the audio and visual signals. He applied the speech to a recognizer that generated a phonetic sequence that was converted to a corresponding viseme sequence using the mapping in Table 2-2. Ashish also stated that he would attempt to extend this work to English language phonemes.

Table 2-2 Faruqui's phoneme-to-viseme mapping rule

Phoneme	Viseme No	Phoneme	Viseme No
a,h	1	g,k,d,n,t,y	7
e,i	2	f,v,w	8
l	3	h,j,s,z	9
r	4	sh,ch	10
o,u	5	th	11
p,b,m	6	Silence	12

Saenko *et al* [29] attempted to use articulatory features to model visual speech. They presented another mapping of English phonemes to 14 visemes as shown in Table 2-3.

Table 2-3 Kate et al. viseme to phoneme mapping rule

Viseme Index	Corresponding Phoneme	Viseme Index	Corresponding Phoneme
1	ax ih iy dx	8	b p
2	ah aa	9	bcl pcl m em
3	ae eh ay ey hh	10	s z epi tcl dcl n en
4	aw uh u wow ao w oy	11	ch jh sh zh
5	el l	12	t d th dh g k
6	er axr r	13	f v
7	Y	14	gcl kcl ng

Saenko's group wanted to design a classifier that would identify some of the phonemes using the four visual features related to lip shape shown in Table 2-4.

Table 2-4 Viseme to feature mapping

Viseme	Lip-Open	Lip-Round
/ao/	Wide	Yes
/ae/	Wide	No
/uw/	Narrow	Yes
/dcl/	Narrow	No

Saenko's group developed an audio-visual database consisting of two speakers and conducted preliminary experiments to classify the phonemes using the features listed in Table 2-4. They obtained classification rates above 85%. The system they developed focused on vowel identification, and the testing was done on two subjects only.

Jintao *et al* [30, 31] worked on visually classifying consonants based on visual physical measures. They developed an audio-visual database consisting of four speakers producing 69 consonant-vowel (CV) syllables. The video recordings from the database were presented to six viewers with average or better lip reading abilities, and visual confusion matrices were obtained. Consonants that were most commonly visually confused with each other were grouped together as one visual unit as shown in Table 2-5

Studies conducted by this group included placing 20 optical markers on the face of a speaker as shown in Figure 2-1. A motion detector designed by Qualisys tracked the 3D positions of these markers. The output of the detector had 51 points for every marker per frame. These points were arranged in matrices, and Euclidean distances between the points in consecutive frames were calculated and used as visual features. These features were then used to train a clustering based classifier using the classes shown in Table 2-5. The recognition accuracy was 38.4% for the spoken CV sequence /Ca/, 36.1% for the spoken CV sequence /Ci/ and 36.0% for the spoken CV sequence /Cu/.

This study is very close to the objectives of our project but it was done only on two speakers. In addition, this study required the use of the 3D coordinates of the points of focus shown in Figure 2-1. This adds some limitations on the applications of this study since 3D coordinates are not available in many applications, such as video conferencing, telephony, and satellite images.

Table 2-5 Jintao phonemic equivalence classes

Visual Unit	Confused Consonants
1	/m,b/p/
2	/f,v/
3	/r/
4	/w/
5	/θ,δ/
6	/ζ,ʃ,d ζ,tʃ/
7	/t,d,s,z/
8	/l,n/
9	/k,g,y,h/

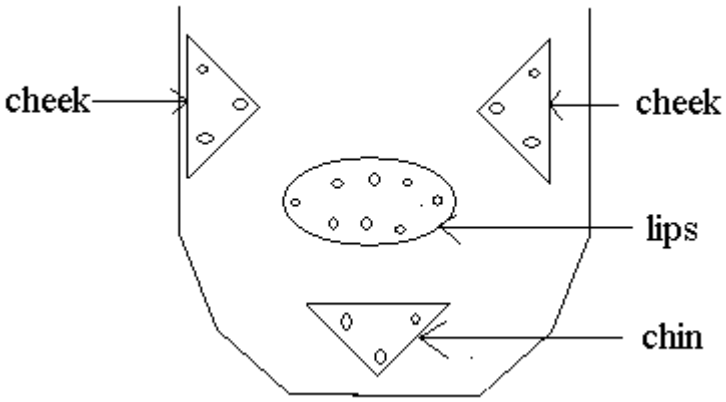

Figure 2-1 Placement of 20 optical markers on speaker face

Warda *et al* [32] attempted to classify visual visemes associated with three French phonemes /ba/, /bi/,/bou/. The group used lip corners and center points of both upper and lower lips as visual features entered into a neural network for the purpose of distinguishing between the video recordings of the three phonemes. The study didn't mention the number of speakers involved. The results are shown in Table 2-6

Table 2-6 Recognition rate of French vowels

	Recognition Rate	
	Training Set	Testing Set
Ba	63.33%	63.64%
bi	73.33%	72.73%
Bou	83.33%	81.82%
Average	73.33%	72.73%

Leszcynski *et al* [33, 34] used three classification algorithms for visemes obtained from the CORPORA database that consists of audio-visual recordings of Polish. The group used two different sets of features to describe the visemes. The first one was based on a normalized triangle covering the mouth area and the color image texture vector indexed by barycentric coordinates. The second procedure performed a 2D Discrete Fourier Transform (DFT) on the rectangular image including the mouth area with respect to small blocks of DFT coefficients. The classifiers in their work were based on Principle Component Analysis and Linear Discriminant Analysis. The group reported that the DFT+LDA exhibits higher recognition rates than MESH+LDA and MESH+PCA methods – 97.6% versus 94.4 and 90.2%, respectively. It is

also much faster than MESH+PCA. The group obtained a 94% recognition rate for the 6 classes shown in Figure 2-2 but the group didn't associate those visemes with corresponding phonemes.

Figure 2-2 Representative images for six major viseme classes

Huang and Chen [35] used Gaussian mixture models (GMM) and Hidden Markov Models (HMM) in mapping an audio parameter set to a visual parameter set in a technique aiming to synthesize mouth movements based on acoustic speech. In this technique, the visual information was represented by lip location (width and height of the outer contour of the mouth), while 13 spectral coefficients were extracted from the acoustical speech representing the audio data. Both audio and visual features were combined to form a single feature vector that was applied to the GMM and HMM. Huang and Chen reported smooth and realistic lip movements with this system. However, the system assumed that both audio and visual information is available, which might not always be the case.

The studies presented in this section presented different applications for viseme–phoneme mappings and applications. The common visual cues in the applications presented in the previous sections were related to motions and positions of the lips.

2.2.2 Visual Perception of Phonemes

Researchers concerned with speech production and lip reading have obtained different viseme-phoneme mappings and identified phonemes that are visually confused with one another. This section will cover the work done in this area.

Kricos [36] conducted a study with 12 female students who had normal hearing with no experience in lip reading. These subjects were presented with black and white video recordings from six different speakers repeating VCV sounds involving 23 English phonemes and 3 vowels. Subjects were asked to identify the consonant being shown. The responses were used to generate confusion matrices for every speaker. Phonemes that were confused with each other for more than 75% of the utterances were grouped together and considered to have the same visual representation (i.e. viseme). Results from this study are shown in Table 2-7

Another study by Benguerel [37] used video recordings of VCV sounds including the consonants /p,t,k,tʃ,f,θ,s,ʃ,w/ and the vowels /i/, /æ/, or /u/. These recordings, obtained from a single female speaker, were presented to 10 subjects. Five of those subjects were hearing impaired, while the remaining five had normal hearing. All subjects were asked to identify the consonant being shown on the video monitor. Consonants that were confused with each other for more than 75% of total utterances were considered to have the same visual representation. Their results are shown in Table 2-8.

Table 2-7 Phonemes visually confused with each other for different speakers (Kricos)

Speaker 1	Speaker 2	Speaker 3	Speaker 4	Speaker 5	Speaker 6
/p,b,m/	/p,b,m/	/p,b,m/	/p,b,m/	/p,b,m/	/p,b,m/
/f,v/	/f,v/	/f,v/	/f,v/	/f,v,s,z/	/w,r, ∂,θ /
/w,r/	/w,r/	/w,r/	/w,r/	/w,r/	/ʃ,ζ,tʃ,d ζ/
/∂,θ/	/∂,θ/	/∂,θ/	/∂,θ/	/ʃ,ζ,tʃ,d ζ/	/t,d,s,z,n,l,j,h/
/ʃ,ζ,tʃ,d ζ/	/ʃ,ζ,tʃ,d ζ/	/ʃ,ζ,tʃ,d ζ/	/ʃ,ζ,tʃ,d ζ/		
/t,d,s,z/	/t,d,s,z/	/k,g/	/t,d,s,z/		
/l/	/l/				
/k,n,j,h/	/k,g,n,j,h/				

Table 2-8 Phonemes visually confused with each other (Benguerel)

Normal Hearing	Hearing Impaired
/p/	/p/
/f/	/f/
/w/	/w/
/ θ/	/ θ/
/tʃ,ʃ/	/tʃ,ʃ/
/t,k,s/	/t,k,s/

Owen and Blazek [18] extended the studies made by Benguerel and Kricos. They used 10 subjects, 5 hearing impaired and 5 normal hearing. All subjects were presented with video recordings for vowel-consonant-vowel (VCV) sequences without sounds. They used 23 English consonants /p,b,m,f,v,t,k,tʃ,f,θ,ð,s,ʃ,w,r,dζ,ζ,d,s,z,g,n,l,h,j/ and 4 vowels /ɑ/, /i/, /u/, and /^/. Subjects were asked to identify the consonant shown on video. The highest overall correct score was 46%. Results from all subjects were gathered in confusion matrices. When a cluster appeared in the matrix, a 75% response criterion was used as a requirement for considering these phonemes to have the same visual representation (i.e. viseme). This criterion resulted in different viseme classes as shown in Table 2-9

Table 2-9 Visemes associated with different vowels (Owen)

Viseme Class	/ɑ/C/ɑ/	/^/C/^/	/i/C/i/	/u/C/u/
Class 1	/p,b,m/	/p,b,m/	/p,b,m/	/p,b,m/
Class 2	/f,v/	/f,v/	/f,v/	/f,v/
Class 3	/θ,ð/	/θ,ð/	/θ,ð/	
Class 4	/w,r/	/w,r/	/w,r/	
Class 5	/tʃ,dζ,ʃ,ζ/	/tʃ,dζ,ʃ,ζ/	/tʃ,dζ,ʃ,ζ/	
Class 6	/k,g,n,l/	/t,d,s,z/	/t,d,s,z/	
Class 7	/h/			

The viseme clustering was consistent between all four vowels for the first two classes. For vowels /ɑ/, /^/, and /i/, the viseme clustering was consistent between all three vowels for the first five classes.

Dodd [38] conducted a review for the literature available on viseme classification and concluded that visemes can generally be classified into nine distinct groups as shown in Table 2-10.

Table 2-10 Dodd's Viseme groups for English consonants

Viseme #	Consonants	Viseme #	Consonants
1	/f,v/	6	/w/
2	/th,dh/	7	/r/
3	/s,z/	8	/g,k,n,t,d,y/
4	/sh,zh/	9	I
5	/p,b,m/		

The studies presented in this section show some of the attempts made in obtaining mappings between phonemes and visemes. Most of the work was based on human response. The consistency of these mappings across different subjects motivated us to investigate if this consistency can be captured by an automatic classifier that is based on visual cues.

A comparison of the results in Tables 2-7 through 2-10 and the results of viseme- to-phoneme mappings in Tables 2-2, 2-3 and 2-5 show that many phonemes are consistently grouped together in one viseme. These common assignments are summarized in Table 2-11

Table 2-11 Common viseme-to-phoneme mapping

Viseme Class	Associated phonemes
1	/p,b.m/
2	/f,v/
3	/θ,∂/
4	/w,r/
5	/tʃ,dζ,ʃ,ζ/
6	/t,d,s,z/
7	/l/

This objective of this study is to classify visemes based on visual cues. The study involves conducting experiments on a set of audio-visual recordings of specific sounds. Since the viseme classes for some phonemes are consistent across different studies, as shown in Table 2-11, a sample representing each class in Table 2-11 will be in the audio-visual data set used for this study.

2.2.3 Lip reading in Speech Recognition

One of the major areas of application for visual cues is the area of audio-visual speech recognition. Several researchers have worked on incorporating lip motion information into systems that were originally based on audio information, hoping to enhance the performance of these systems. This section covers some of the work done in this area.

In a study of audio-visual signal processing, Chen [13] stated that an automatic lip reading system involves three core components: visual feature extraction, audio feature extraction, and the recognizer or classifier. The automatic lip reading system proposed by Chen used both visual and audio information. Lip movements were used as the visual features. The audio signal was divided into frames and converted into sixteenth-order linear prediction coding (LPC) coefficients to create the audio features. The audio and visual features were combined in one vector and applied to a hidden Markov model (HMM) for final recognition.

For the lip-tracking phase, Chen modeled the color distribution of the face pixels and of the background and then used a Gaussian function to extract the face of the speaker. After extracting the face, a template resembling the shape of the lips was used to extract the corners and height of both the upper and lower lips. This process was repeated for every image frame. One of the limitations of this technique was that the speaker needed to be in front of the camera. Chen compared the performance of a HMM based speech recognizer with audio only input, audio-visual input and visual only input. The audio signals were corrupted with additive white Gaussian noise at various SNRs ranging from 32 dB to almost 16 dB.

The study showed that at an SNR of 16dB the recognition rate of the audio-visual based system was almost four times higher than the recognition rate of the audio-based system. The differences between the recognition rates became less as the SNR increased, but the audio-visual

system consistently had higher values. At an SNR of 32 dB, both systems performed at about the same rate. Chen concluded that automatic lip reading could enhance the reliability of speech recognition. He added that through lip synchronization with acoustics, it would be possible to render realistic talking heads with lip movements synchronized with the voice.

Luettin and Dupont [39] stated that the main approaches for extracting visual speech information from image sequences can be grouped into image-based, geometric-feature-based, visual-motion-based, and model-based approaches. In the image-based approach the gray-level representing the mouth is either used directly or after some preprocessing as a feature vector. The visual-motion-based method assumes that visual motion during speech production contains relevant speech information such as lip movement. The geometric-feature based approach assumes that certain measures such as the width or height of the mouth openings are important features. In the model-based approach, a model for the visible speech articulators, usually lip contours, is built and is described by a small set of parameters. The group developed a large vocabulary continuous, audio-visual speech recognizer for Dutch using different representations of visual cues and showed that a combined audio-visual recognizing system improves upon audio-only recognition in the presence of noise.

Petjan [40] also developed an audio-visual speech recognizer that used lip height and width as visual cues applied with the acoustic waveform to the recognizer. Petjan's results confirmed Chen's claim of obtaining higher recognition rates with the addition of visual cues.

Some researchers have used the image of the entire mouth area as a visual feature applied to a speech recognizer together with audio cues [41, 42]. Li et al.[42] used eigen vector analysis in lip reading. In the training part of their approach, they formed a vector consisting of all gray level values of pixels representing the mouth in all frames of a sequence representing one spoken

letter from the English alphabet. Next they formed a training matrix containing several such vectors and computed Eigen vectors for each letter in the alphabet. In the classification stage, a sequence representing an unknown letter was projected on the model of Eigen space for each letter, and a projection close to "1" represented a match. Li and his group applied their technique to ten spoken letters [A-J] using one person only. The success rate of recognition varied from 90-100% depending on the letter to be recognized.

Potaminos and Chalapathy [43] investigated the use of visual, mouth-region information in improving automatic recognition of the speech. The visual information in the system was represented by the highest 24 coefficients from a Discrete Cosine Transform (DCT) of a 64x64 pixel region-of-interest containing the speaker's mouth. The audio part of the signal consisted of 24 ceptral coefficients of the acoustical speech signal. Both features were combined in a vector that was then applied to a HMM. Incorporating the visual information improved the SNR by 61% over audio only processing.

Another popular visual feature used in speech recognition was based on visual cues from the lips and jaw movements. Paul et al.[15] designed a speech recognizer system that incorporated lip reading information with the acoustic signal to improve speech recognition. The image of the face was supplied to a neural network that extracted the mouth corners and lip movement information. This extracted information was then applied with the acoustic information to a Multi State Time Delay Neural Network (MS-TDNN) to perform the final recognition. Paul and his team stated that compared to audio-alone recognition, the combined audio-visual system achieved a 20-50% error rate reduction for various signal/noise conditions.

Baig and Gilles [44] presented a new neural architecture, called a spatio-temporal neural network (STNN) and used it in visual speech recognition. Biag and his group chose four points

on the lips and generated a time signal by tracking these four points in successive image frames. These time signals together with the acoustical signal were used as inputs to the STNN for recognition purposes. They tested their system on 510 audio and visual sequences of numbers spoken in French. They used 260 sequences for training the network, while the remaining 250 sequences were used for testing the performance. Although the test and training samples were from the same person, their results showed a success rate of 77.6%. The study did not compare the performance of the system for audio only input.

Luettin and Dupont [39] combined the inner and outer lip contour positions together with the lip intensity information and formed a vector of 24 elements representing the visual features. The audio features were obtained by choosing 24 linear prediction coefficients. The audio and visual features were applied to a HMM for speech recognition. They tested their system on clean speech and reported an error rate of 48% with visual features only and 3.4% with the audio signal only. When both audio and visual features were used, the error dropped to 2.6%.

Goldschen [45] applied Hidden Markov Models (HMM) as classifiers in a speech recognizer having both audio and visual input data. He also studied which features led to better speech classification decisions. The feature set he preferred was associated with the dominant mouth movement in terms of upper and lower lips, rather than the lip positions. Mase and Pentland [46] reached the same conclusion.

The examples shown in this section explored many speech recognizers that utilized both audio and visual information. Performance of speech recognizers improved when visual cues were included in the system. Visual features extracted from the lips were commonly used in most applications, which is consistent with the studies previously discussed in Sections 2.1 and 2.2.

2.3 MISCELLANEOUS APPLICATIONS OF AUDIO-VISUAL SIGNAL PROCESSING

Human beings can use the visual information available in lip movements or facial expressions to separate two sounds coming from two different sources. Okuno et al. [47] attempted to design a system that would use visual information in enhancing sound source separation. They used the movement of the mouth as an indication of a sound source. This information was sent to a module that checked whether the image information and the sound information are from the same source. If it is, the position information is recorded. Otherwise, the visual module is moved to focus on another sound source. Okuno stated that adding the visual information increased the dimension of the problem. However it provided an accuracy of a few degrees for a point source at 2 meters distance, which was higher accuracy than audio only sound source separation.

Speaker detection is another area in which audio-visual information has been combined. Speaker detection is a useful tool in video conferencing, where a camera needs to focus on a person identified as the speaker. Cutler et al.[48] proposed a measurement for the correlation between the mouth movements and the speech and used it in a time delayed neural network (TDNN) to search for a certain speaker. The system was able to successfully locate a single speaker. Their results did not provide a quantitative measure for the accuracy of the device with and without the visual information.

Another area in which visual information has been used is sound sensing. Takahashi and Yamazaki[49] proposed a sound sensing system that used audio-visual information. The system was divided into two subsystems: an audio subsystem and a visual subsystem. The audio subsystem extracted a target signal with a digital filter composed of tapped delay lines and

adjustable weights. These weights were modified by a special adaptive algorithm called "cue signal method". For the purpose of adaptation, the cue signal algorithm needed only a narrow bandwidth signal that is correlated with the power level of the target signal. This narrow band signal was called the “cue” and was generated from the visual subsystem.

The authors stated that fluctuations in sound power due to lip movement correspond to a visual stimulus in the image. Therefore, by locating the lips in an image, the system obtained a visual estimate of the sound power by the squared absolute time difference in successive image frames. Another estimate for the sound power was obtained directly from the audio signal. Both audio and visual estimates of the sound power were multiplied together to form the cue signal. With the visual cues used, their results showed a 96% improvement in object localizing over audio only based object localization.

Facial expressions are another area of focus of researchers, since many people can understand emotions based on the facial expressions appearing on the people surrounding us [50]. Craig et al. [51] stated that facial expressions can indicate pain. Katsikitis and Pilowsky[52] mentioned that facial expressions reveal brain functions and pathology.

Ekman and Friesen [53] developed an anatomically based Facial Action Coding System (FACS), and many researchers [54-56] working in this area mention that FACS is the most comprehensive method of coding facial displays. This coding system was obtained by viewing videotapes in slow motion of a large sample of recorded facial expressions and then coding those expressions to form action units. The FACS contained more than 7000 facial expressions. In a later study [54], Ekman and Friesen proposed that emotion codes can be obtained by specific combinations of FACS action (i.e. fear, joy, sadness, anger, disgust and surprise). Hegely and

Nagel [56] collected these emotions together to form the Emotions Face Action Coding System (EMFACS).

Izard [57] developed another anatomically based systems which requires slow motion viewing of videotapes. He called it the Maximally Discriminate Facial Movement Action Coding System (MAX). Compared with FACS, MAX is less comprehensive, and is intended only to code emotion based facial displays, while FACS is intended for displays that are not only emotion related.

Essa and Pentland [58], Mase and Pentland [59], and Yacoob and Davis [60] attempted to use optical-flow-based approaches to discriminate facial displays (e.g. fear, joy, surprise). Such approaches were based on the assumption that muscle contraction causes skin deformation. This skin deformation changes the optical spectrum appearing on the face of the speaker. In a digitized image sequence, algorithms for optical flow extract motion from the texture changes in the skin, and the pattern of this motion can be used to discriminate facial displays.

Pantic [61] and his group designed an expert system they called Integrated System for Facial Expression Recognition (ISFER). This system performs recognition and emotional classification of human facial expressions from a still, full-face image. At the time of publishing their work, the system was capable of automatically classifying face actions into six emotion categories (happiness, anger, surprise, fear, disgust and sadness).

This discussion affirms the claims that visual cues may contribute to speech perception. However, the applications of facial expressions were limited to identifying emotions and not speech. Despite this limitation, identifying emotions may help in speech perception, since different emotions involve the use of different vocabulary.

2.4 SUMMARY

Despite the wide range of audio-visual applications reviewed in this chapter, there is still little work done in the area of automatic recognition of visemes based only on visual cues. The work done on recognizing visemes was based on the visual response of subjects as detailed in Section 2.2.2. The consistency of viseme grouping in speech psychology test results shown in Table 2-11 motivated this attempt to design an automated viseme classifier that is based on a set of visual cues only. The outlines of this work were detailed in Figure 1-2.

There are many factors to be considered in choosing which phonemes to focus on in developing the audio-visual data needed for this study. Since most audio-visual applications utilize 2-D imaging devices, this study shall focus on visual features that could be extracted from 2-D images of speakers. The recording device used in this study was a Motion Analysis system (ExpertVision, Inc) VP110 that traces the 2-D motion of optical reflectors placed anywhere on the face of the subject. The VP110 is manufactured by the Motion Analysis Corporation located in Santa Rosa, California. The optical reflectors used with the device had a circular shape with a reflector side and an adhesive side that sticks to the point of interest. This limits the ability to distinguish phonemes that involve the inside part of the mouth such as sounds within classes 5 and 7 in Table 2-11. The voiced consonants in English are /b/ /d/ /g/ /v/ /ð/ /n/ /l/ /w/ /j/. In addition, the English sounds that involve the lips, jaw and teeth in production are /b,p,m,f,v,th,w,/. One sound from each of the classes of the common viseme-to-phoneme mapping (Table 2-11) was chosen to represent the class to be distinguished. These phonemes are shown in Table 2-12.

Table 2-12 VCV sounds to be rerecorded

Viseme Class	VCV sound
Class 1	/b/
Class 2	/v/
Class 3	/ð/ => the
Class 4	/w/
Class 5	/d/, /z/

The remaining two classes of Table 2-11 were not studied since they are sounds produced inside the mouth. Phonemes /d/ and /z/ were chosen together from the same class to test if the 1st and 2nd derivatives of the lip motion waveform can capture the difference between both sounds, even though subjects could not distinguish them consistently.

The common viseme-to-phoneme mappings showed that the viseme classes did not change when vowels /ɑ,i,^/ were presented to subjects in association with different consonants. In order to test whether the classifier can distinguish between vowels, designs and tests on the VCV recordings were implemented using VCV words with both vowels /ɑ,i/. This resulted in a total of 12 VCV sequences where the consonants are shown in Table 2-12 and the vowels are /ɑ,i/. The 2nd vowel in the VCV sequence was emphasized during the sound production.

The correlation between the acoustics and visual cues discussed in Section 2.1 showed that lip motion has high correlation with the acoustic signal being produced. In addition, work on audio-visual speech recognition showed that visual features related to lip motion were the most popular ones to use. The visual features representing the first block of Figure 2-1 were extracted

from the 2-D images of lip motion. The study assumes that the location of lips is already determined, and it will not consider techniques to extract lips from recorded video sequences.

3.0 EXPERIMENTAL METHOD

This chapter explains the experimental setup employed in the study. The first section presents the data recording method. Section 3.2 explains the pre-processing applied to waveforms. Section 3.3 discusses the feature extraction. The final section describes the linear discriminant analysis method.

3.1 DATA RECORDING

3.1.1 VP110 Motion Analyzer

The VP110 Motion Analysis System (ExpertVision, Inc) analyzes the motion of objects within single or multiple video frames. The system features a real-time data acquisition system together with data analysis capabilities. The ExpertVision system consists of the following units: infra-red optical reflectors, video camera, array of infra-red LED lights, computer, and the VP110 unit [62].

Figure 3-1 shows how a subject with five optical markers placed around the subject's face sits directly in front of the camera of the Analyzer. An array of infra-red LEDs is attached to the camera to insure that the amount of infra-red light reflected off the optical markers is higher than the amount reflected off the remaining parts of the face. The Motion Analyzer has a

threshold for the brightness of the infra-red light received from the camera. This threshold ensures that only objects with high brightness due to infra-red light reflections are detected by the camera. Once the recording process starts, the Motion Analyzer tracks the optical reflectors at a frame rate of 60 Hz and identifies their outlines in consecutive frames. These outlines are used as input to ExpertVision software to identify the x-y coordinates for the center point of individual reflectors in each frame. A microphone, used to record the audio signal spoken by subjects, is sampled at 22050 Hz by the computer. The recorded audio signal was not used in the study.

The ExpertVision software processes the recorded video sequence frame-by-frame, calculates the coordinates of the center of each optical reflector based on the coordinate system shown in Figure 3-2, and stores the coordinates in a file called the centroid file. The centroid files are processed further in Matlab to generate a path associated with each centroid in consecutive frames. The path files are used to generate distance waveforms. Then, the visual features are extracted. This process is discussed further in Section 3-3.

3.1.2 The Recording Procedure

Five reflectors were placed on the face of the subjects as shown in Figure 3-2. The motion of these reflectors represented the visual features associated with the VCV sequence as discussed in Section 3.3. The reflectors were placed on the mid points of upper and lower lips, lip corners and the upper nose bridge. The upper nose bridge is a point on the face that does not move while a subject is speaking. It was used as a correction factor for the effect of head movement during the recordings. Participants sitting in front of the LED lights repeated the desired VCV sequence. The recording of the audio-visual sounds involved the following steps:

- Participants' hearing was screened using the Welch Allyn AudioScope 3 Screening Audiometer. This instrument provides the means for quickly checking that the ear canal is patent and that the subject can hear tones presented at 500, 1000, 2000, and 4000 Hz at 25 dB hearing level. These are frequencies that are of primary importance for hearing speech and to assure accurate hearing of speech at conversational levels. Participants were required to detect all four test tones, at least in one ear, to be included in the study.
- The required VCV word was presented to participants via a head-phone, thus guiding them in producing the proper clear VCV sequence. The guide signal was recorded by a PhD audiology student capable of producing clear VCV sequence. The guide signal emphasized the 2^{nd} vowel in the sequence.
- Participants repeated the VCV word they heard in the head phone several times closing their lips after each repetition.
- Participants repeated the desired VCV words while sitting directly in front of the camera.
- The recording time for each VCV sequence was 45 seconds.
- Once the recording of a certain VCV word was done, the collected data was saved and the recording of the next VCV word began.

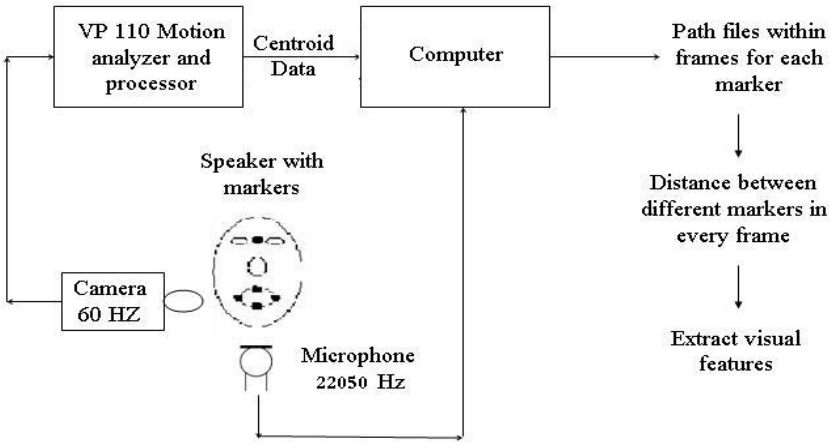

Figure 3-1 Experimental setup for audio-visual data recording

Figure 3-2 Location of optical reflectors and coordinate system for tracking of lip movements

3.1.3 The Recorded Audio-Visual Data

Audio-visual recordings from 28 participants (26 females and 2 males) were obtained. The participants were monolingual speakers of Native American English and capable of producing clear speech. The protocol for the recordings was approved by the University of Pittsburgh Institutional Review Board (IRB).

Table 3-1 lists the number of word utterances obtained for each VCV sequence for all 28 speakers after the processing of centroid and path files was completed. Some speakers produced more utterances than others. In addition, the Motion Analyzer sometimes failed to capture some of the optical reflectors in consecutive frames, reducing the number of available word utterances. For speakers 4, 10, 11, 19 and 21 multiple reflectors were dropped out in consecutive frames for several seconds. This led to discarding the recordings of some of the VCV sequences associated with those speakers. The brightness threshold set in the Motion Analyzer was not able to remove other bright spots on the face of some participants. Teeth and even cheeks of some participants resembled optical reflectors and were captured by the motion analyzer. This led to additional centroids appearing in each frame, causing confusions in the path assignments in consecutive frames.

A criterion of a minimum number of 8 word utterances was used to include a speaker in the analysis. Exception was made for speakers 1, 7 and 14 who had 8 or above word utterances in 11 of the VCV sequences and had 7 utterances in the 12^{th} VCV word. The missing utterance for each of these speakers was compensated by adding the average of the other 7 utterances. This resulted in including audio-visual data from the 18 speakers shown in bold fonts in Table 3-1. Audio-visual data coming from the remaining 10 participants was not used in the study.

Table 3-1 Number of utterances for each word by all speakers

#	Ab	Av	At	Aw	Az	Ad	Ib	Iv	It	Iw	Iz	Id	Total
	1	2	3	4	5	6	7	8	9	10	11	12	
1.	**11**	**12**	**12**	**12**	**12**	**12**	**9**	**10**	**12**	**7**	**13**	**12**	**134**
2.	**14**	**12**	**11**	**9**	**10**	**9**	**9**	**8**	**10**	**10**	**11**	**10**	**123**
3.	**10**	**10**	**9**	**10**	**10**	**11**	**8**	**10**	**8**	**9**	**10**	**11**	**116**
4.	0	16	10	14	15	14	8	8	13	14	12	10	134
5.	**12**	**12**	**12**	**11**	**8**	**8**	**11**	**11**	**11**	**12**	**12**	**11**	**131**
6.	**11**	**11**	**11**	**8**	**10**	**14**	**12**	**15**	**11**	**13**	**14**	**14**	**144**
7.	**10**	**9**	**8**	**11**	**11**	**11**	**7**	**10**	**10**	**10**	**8**	**9**	**114**
8.	**13**	**15**	**17**	**16**	**17**	**11**	**17**	**17**	**16**	**17**	**16**	**12**	**184**
9.	**10**	**12**	**13**	**13**	**12**	**11**	**12**	**11**	**11**	**9**	**8**	**10**	**132**
10.	12	11	11	10	6	8	10	0	10	10	10	10	108
11.	0	14	6	6	8	7	9	6	9	9	7	8	89
12	**9**	**9**	**10**	**10**	**10**	**10**	**11**	**8**	**10**	**11**	**10**	**11**	**119**
13	6	8	8	8	10	9	11	10	10	8	11	8	107
14.	**12**	**12**	**10**	**13**	**13**	**7**	**13**	**13**	**13**	**11**	**12**	**10**	**139**
15.	**12**	**13**	**13**	**13**	**13**	**13**	**13**	**14**	**13**	**13**	**12**	**11**	**153**
16.	10	9	9	7	8	8	8	12	7	12	8	8	107
17.	**12**	**9**	**12**	**12**	**10**	**10**	**9**	**10**	**11**	**11**	**9**	**8**	**123**
18.	**8**	**12**	**14**	**18**	**17**	**14**	**13**	**10**	**13**	**12**	**8**	**12**	**151**
19.	19	15	16	0	0	14	11	8	11	14	13	13	134
20.	**9**	**12**	**10**	**9**	**10**	**12**	**11**	**12**	**11**	**10**	**9**	**8**	**123**
21.	7	0	12	12	9	7	9	12	8	9	10	12	107
22.	**11**	**14**	**15**	**8**	**15**	**15**	**16**	**15**	**10**	**13**	**15**	**15**	**162**
23.	**12**	**11**	**11**	**12**	**10**	**11**	**12**	**10**	**10**	**11**	**10**	**11**	**131**
24.	8	9	9	10	8	9	7	7	8	9	8	8	100
25.	**11**	**11**	**11**	**14**	**14**	**15**	**13**	**12**	**9**	**12**	**13**	**13**	**148**
26.	**15**	**15**	**14**	**12**	**13**	**9**	**12**	**14**	**13**	**11**	**11**	**12**	**151**
27.	6	9	9	9	10	11	13	10	12	10	9	10	119
28.	10	9	9	11	11	9	9	8	8	8	6	5	103
Total	282	313	315	303	305	305	310	299	307	315	306	304	3664

3.2 PRE-PROCESSING THE WAVEFORMS

The recording system produces two waveforms per reflector. Each waveform shows the position of the reflector on one axis (x or y) in every frame as the subjects repeats a VCV word. Thus, the total number of waveforms to be processed for every VCV word is ten. Figure 3-3 shows an example of the waveforms associated with the lip markers for one subject repeating /aba/ 8 times. These waveforms are generated by the ExpertVision system and stored as centroid files on the computer.

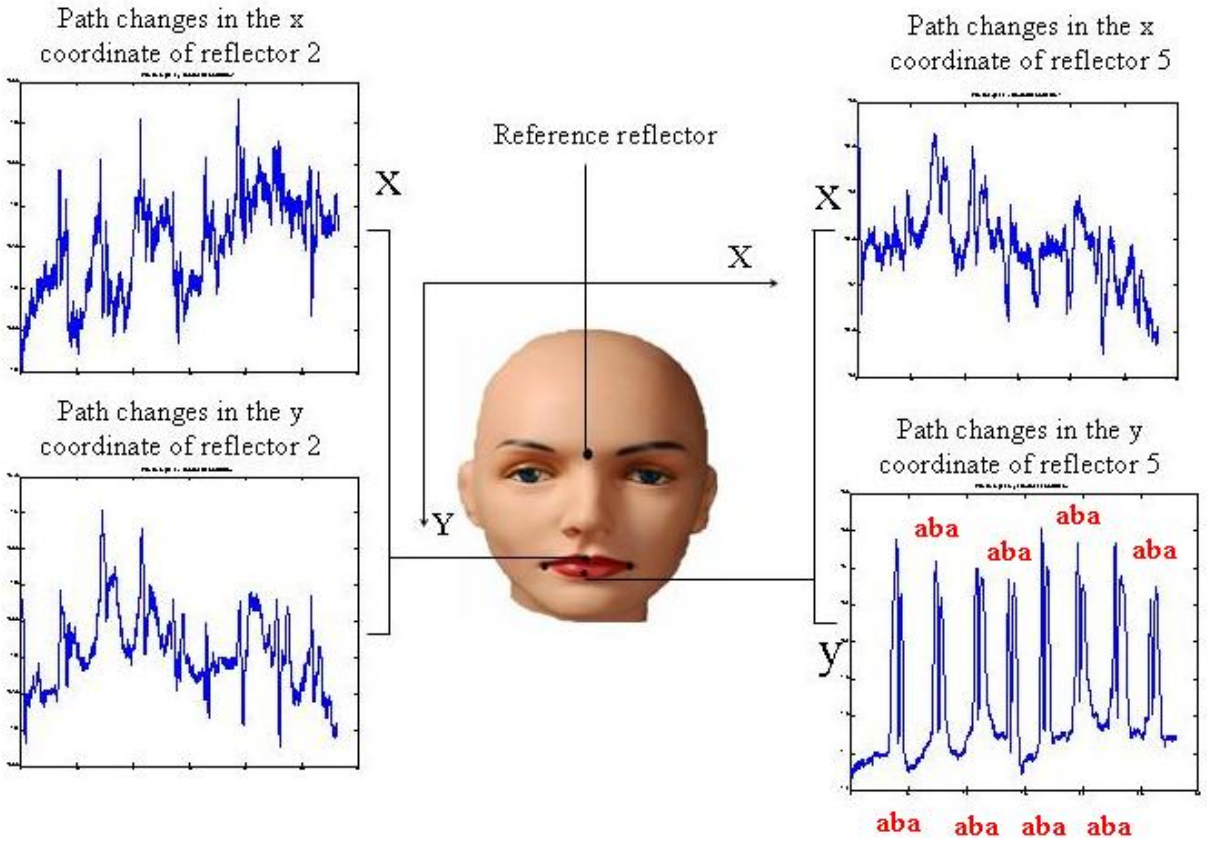

Figure 3-3 Waveforms associated with each reflector

3.2.1 Generating Path files from Centroid files

In every frame of the image, the Motion Analyzer scans the image from top to bottom and left to right. When a reflector is detected, a number is assigned to it and the x-y coordinates for that reflector are recorded. There is no guarantee that the same number assignment is given to every centroid in consecutive frames. A certain centroid might be assigned number 2 in one frame, but in the following frame, it may be assigned as number 3. This confusion increased when the reflectors within a frame were along the same line in the x-direction.

The assignment confusion is corrected by scanning the generated centroid files frame by frame to ensure that every reflector has the same assignment in consecutive frames. This process involved developing a Matlab code to perform that following tasks:

- Check the first frame and assign a path number for the x and y coordinates of each centroid as shown in Table 3-1

Table 3-2 Path assignment in the first frame of the centroid file

	Path 1 Reference Path	
	Path 2 Upper lip	
Path 3 Left lip corner		Path 4 Right Lip corner
	Path 5 Lower Lip	

- Check the next frame f_j, $1 < j <= n$, where n is the total number of frames:
  - Use the last coordinate location P_i for Path i (i=1,2,3,4,5) from previous frame f_{j-1}
  - Measure Euclidean distance between all five centroids in f_j and P_i and assign the centroids in f_j with the closest centroid in path P_i
  - Go to the next path P_{i+1} in frame f_{j-1} and measure the Euclidean distance between the remaining centroids in f_j and P_{i+1} and assign the centroid with the shortest distance in f_j with the path P_{i+1}.
  - Repeat the above steps for the remaining centroids in frame f_j
  - Go to the next frame
- A flag is used to determine if all five centroids were detected in a frame. If one centroid is dropped out in a frame, the path associated with that centroid is marked and a search is made for the closest centroid to the missing path in future frames. The algorithm then linearly interpolates points across missing frames to keep the path connected between frames.
- If a certain path P_j is dropped for more than 60 frames, that path is dropped during those 60 frames while preserving the other paths. P_j continues to be tracked when it appears at later frames.
- If the number of detected centroids in a given frame f_j is greater than five, the Euclidean distance between them and the five centroids detected in the previous frame f_{j-1} is calculated. Each centroid in f_j is assigned to the closest centroid in f_{j-1} and the extra centroids from frame f_j are discarded
- If more than one centroid is missing in a given frame, the frame is dropped.

The corrected assignments are stored in path files. There are 10 path files marking the x-y coordinates for each reflector in consecutive frames.

3.2.2 Removing the effect of head motion from every frame

The reference reflector location (at the upper nose bridge) was used to track the location of the head in every frame. Since the visual features to be extracted rely on the x-y coordinates of the reflectors in consecutive frames, the effect of head motion on the changes in these coordinates should be minimized. This correction was done by the following technique:

- Calculate the average location of the reference reflector in x and y coordinates in the first 60 frames, and use it as the reference point in that window.
- Find the difference between the above reference point and the actual location of the reflector in each frame within the 60 frames window in both x and y directions.
- Add the difference in x-direction as well as the difference in y-direction to the location of the four reflectors around the lips in all the frames within the window.
- Move to the next 60 frames window

This correction reduces the effects of head motion in the x-y directions that is parallel to the face of the camera lens. It does not compensate perfectly for forward or backward face tilting. To assess the error that could be introduced by tilting, the actual motion of the forehead reference centroid was monitored across speakers in consecutive frames. A sample waveform for this motion for one VCV sequence from one subject is presented in Figure 3-4. The motion for the forehead reference point in consecutive frames was between half and one pixel. The range of the overall motion for the forehead reference point in different speakers was between 1-2 pixels which is close to the pixel noise level (0.5 pixels) in the Motion Analyzer system itself [62]. This

indicates that the subjects had minimal head motion during the 45 sec length of a recording session, and error due to head tilting is not significant.

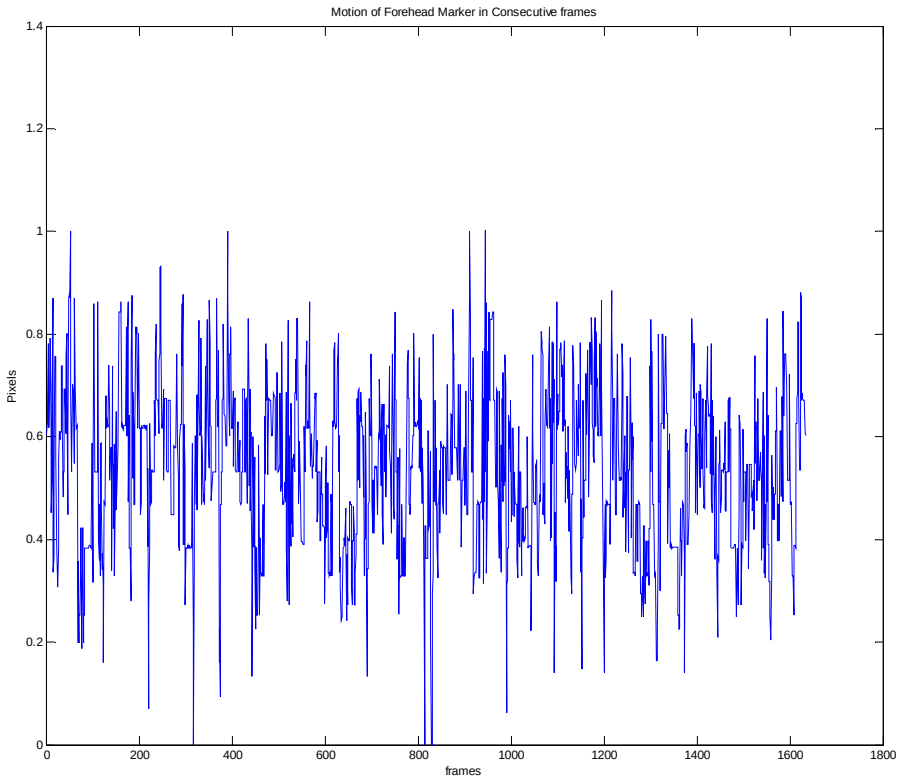

Figure 3-4 Motion of forehead reflector in consecutive frames

3.2.3 Generate the distance waveforms

It is usually desirable to simplify a problem by reducing the number of variables without jeopardizing the information stored in those variables. The number of variables in this problem

can be reduced by decreasing the number of waveforms to be processed. This can be achieved by converting the path files into distance waveforms. The following distance waveforms were generated:

1. Euclidean distance between upper and lower lips in consecutive frames
2. Euclidean distance between lip corners in consecutive frames
3. Euclidean distance between upper lip and forehead reflector in consecutive frames.
4. Euclidean distance between lower lip and forehead reflector in consecutive frames.

The distance between the upper and lower lip waveforms superimposed on the audio file associated with the utterance for the eight repetitions of the word /ɑbɑ/ for one subject is shown in Figure 3-5. The peaks of the waveform correspond to the distance between upper and lower lips while speaking. The first peak is associated with producing the sounds /ɑ/, the minimum is associated with producing the consonant /b/, and the second peak represents the mouth opening while producing the second vowel /ɑ/ in the sequence.

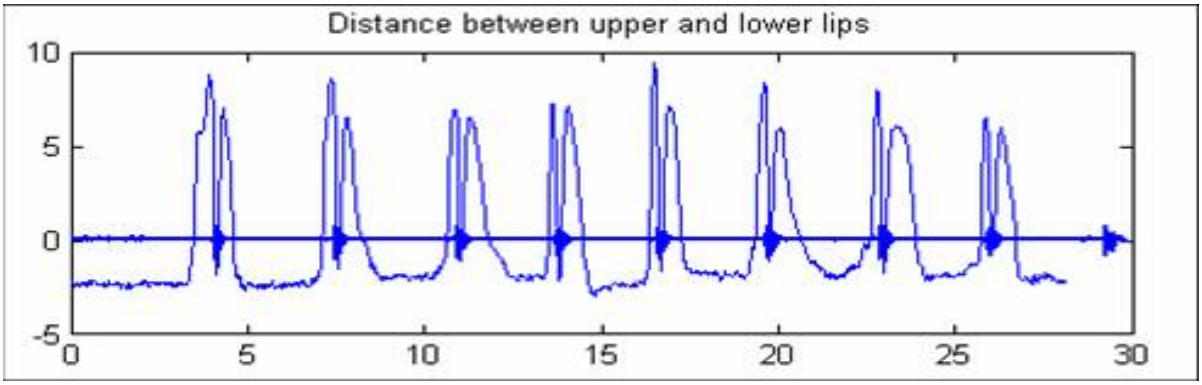

Figure 3-5 Upper/Lower distance waveform for "aba" with the audio signal superimposed

The focus on distance waveforms reduced the number of waveforms to consider. In addition, it further reduced the effect of any possible head motion remaining in the path files, since the distance between two points is independent of the location of the head as long as head tilting is not involved. Figures 3-6 through 3-9 shows examples of these waveforms extracted from four different VCV sequences repeated by the same speaker. The x-axis in these figures represents the frame number while the y-axis represents the amplitude of the distance waveform.

There is a pattern repeating in every waveform around the extremas of the waveform. The objective now is to quantify this pattern by using a set of parameters (visual features) to characterize these waveforms. In addition, it is noted that there is a correlation between the 1st waveform representing the distance between upper and lower lips and the 4th waveforms representing the distance between the lower lips and the forehead reference point. This correlation shows that the role played by the upper lips in producing a word is much smaller than the role played by the lower lips and that the lower lips dominate the upper/lower distance waveform. In the remaining analysis of the data, only the upper-lower lip waveform will be used, and the lower-lip waveform will discarded.

Figure 3-6 and Figure 3-7 show the same speaker repeating two different words. Figure 3-6 shows that the speaker had 12 utterances of the word /ɑbɑ/, and Figure 2-7 shows 19 utterances of the word /ɑðɑ/. In addition to speaking at different rates, the amplitudes of the waveforms differ from one speaker to another. This demonstrates the need for amplitude as well as time normalization before visual features can be extracted to represent a specific word.

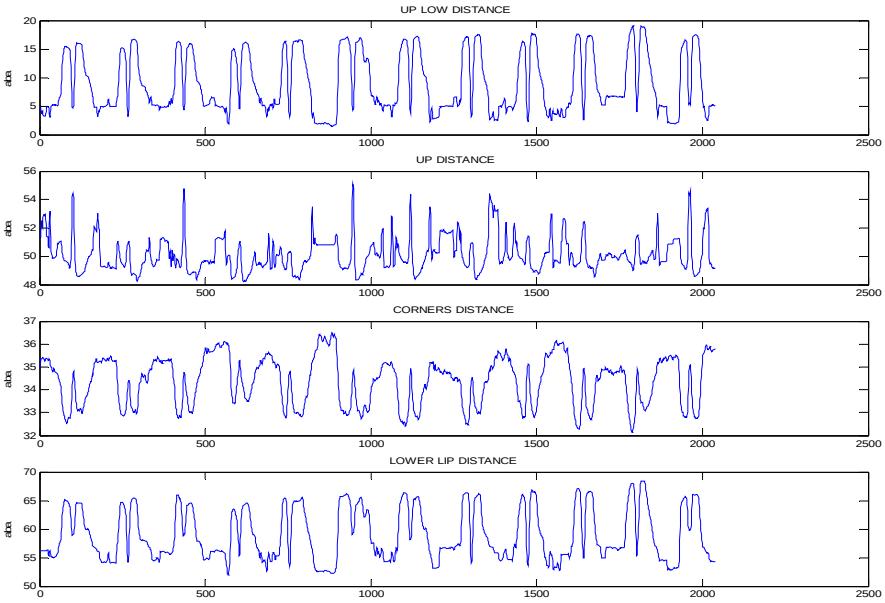

Figure 3-6 Four distance waveforms associated with the VCV word "ɑbɑ"

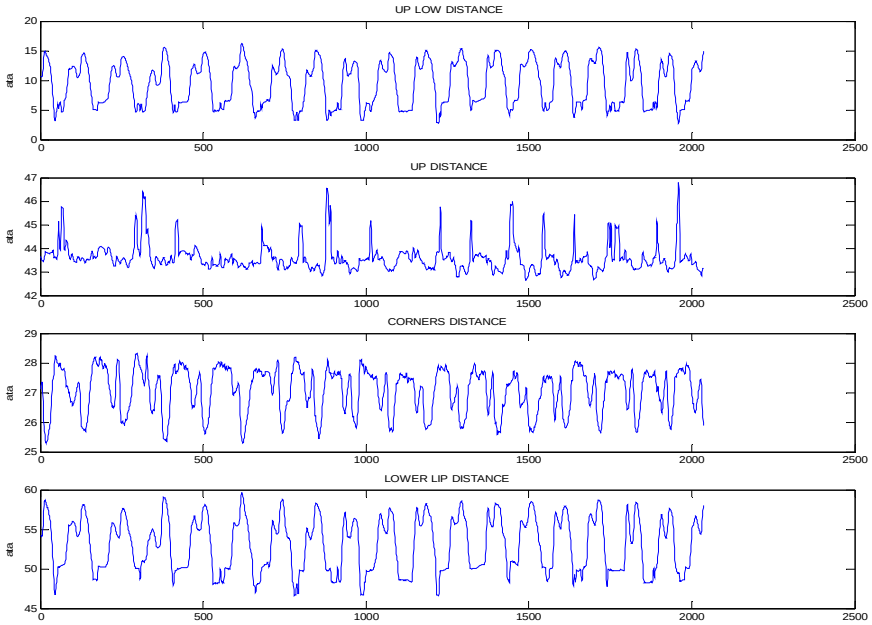

Figure 3-7 Four distance waveforms associated with the VCV word "ɑðɑ"

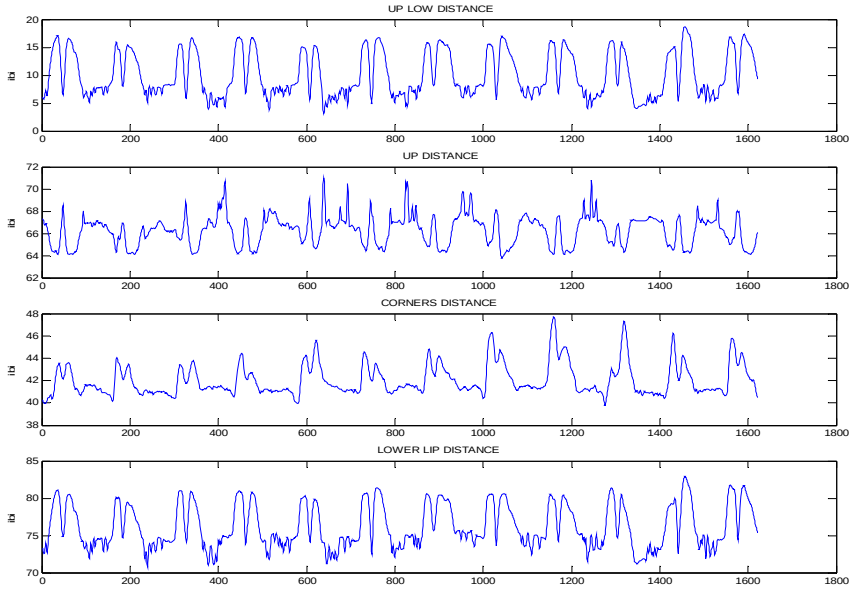

Figure 3-8 Four distance waveforms associated with the VCV word "iði"

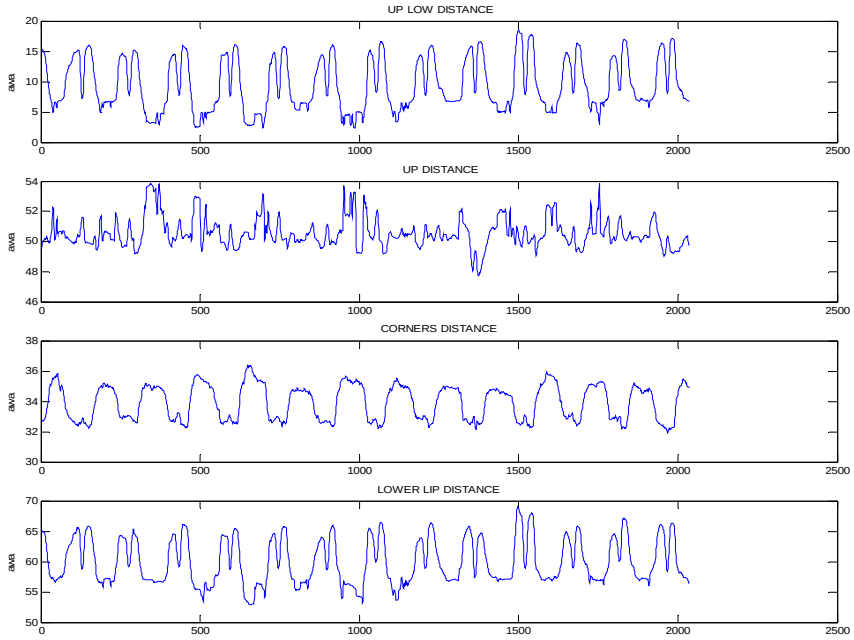

Figure 3-9 Four distance waveforms associated with the VCV word "ɑwɑ"

3.2.4 Obtaining Single Utterances

The distance waveforms in Figures 3-6 through Figure 3-9 showed the need for applying time-and-amplitude normalization to the waveforms to compensate for speaking at different rates. But before the time-and-amplitude normalization can be obtained, distances associated with single utterances from every word need to be obtained for every speaker. The following steps were used to obtain single utterances for every VCV word:

- Start with the upper-lower lips as well as lip-corners distance waveforms calculated earlier.
- Use a threshold to determine the starting point of an utterance.
- The algorithm displayed the distance waveforms and requested the user to choose the word length as well as the number of points to go behind the starting point identified in the previous step in each waveform.
- The algorithm allowed five possible word lengths options 120, 108, 96, 84, and 60 samples. VCV words repeated at a fast rate require shorter word length, while VCV word sequences repeated at slower rates require longer word lengths. If the selected word length is less than 120, then zeros are added before and after the word so that the word is stored in the middle of the record.
- Samples from the starting point determined in step 2 to the ending point that depends on the length of the word are stored in a vector of 120 samples (i.e. 2 sec).
- The resulting processed words are displayed. The algorithm waits for an input from the user to verify that the word length chosen was long enough to capture each complete word utterance. If more than one word utterance was captured or if the samples captured were not enough to obtain one complete utterance, then the user can

go back to step 3 to modify the word length. If the parameters worked, then files are stored.

An example for the results of breaking up the words into single utterances is shown in Figure 3-10 for two distance waveforms. The ninth waveform in every figure shows the average of all utterances together plus and minus a standard deviation about the average. The waveforms show good consistency for the word spoken by the same speaker, which was common across speakers.

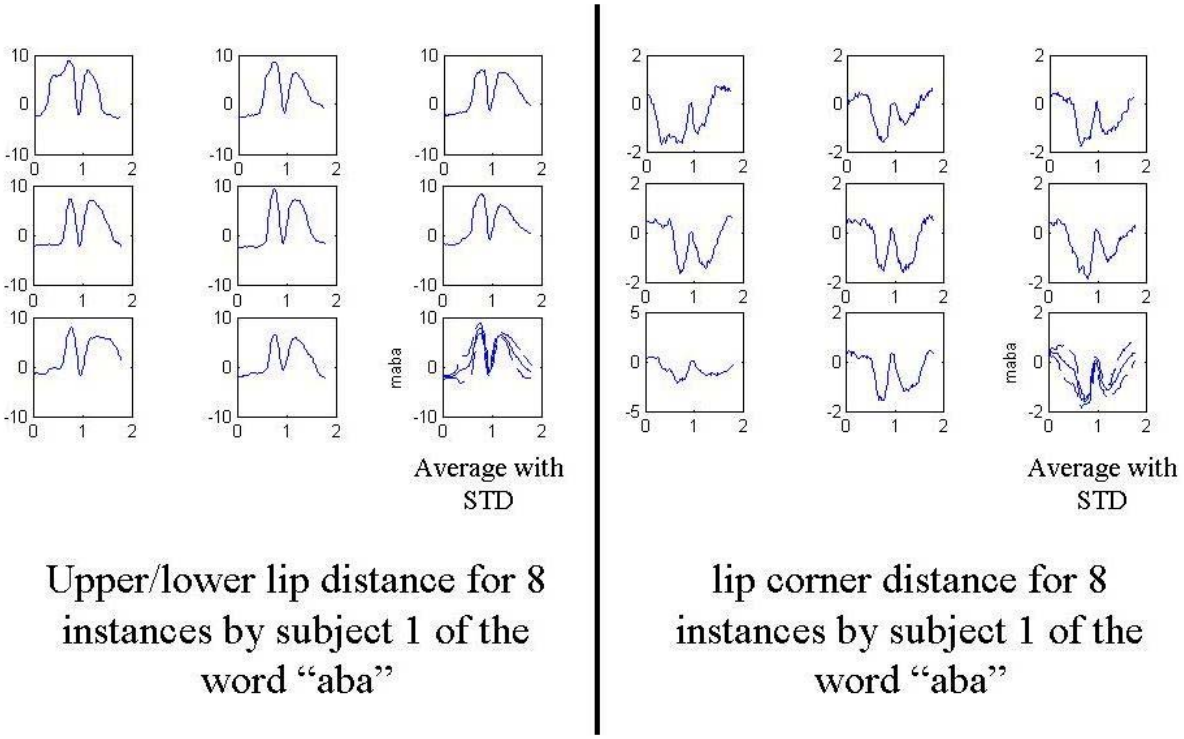

Figure 3-10 Broken utterances for the word /ɑbɑ/ together with the mean for each speaker

3.2.5 Time-And-Amplitude Normalization

The objective of time–and–amplitude normalization is to have the same time and amplitude scales for all word utterances before extracting the visual features. Subtracting the mean of the word utterance ensures that all utterances have zero means.

Several normalization schemes were evaluated. One method for amplitude normalization is to divide each of the utterances by the maximum amplitude of the absolute value of the waveform. This forces all utterances to have maximum amplitude of one. Another method popular in amplitude normalization involves dividing each utterance by the standard deviation present in the waveform representing the utterance. The visual features used in the study and discussed in Section 3.3. were related to the extremas present in the distance waveforms. Dividing by the maximum of the absolute value of the waveform forced that feature to have an amplitude of one across utterances of all speakers. This removes the variability in one of those extremas, which may result in losing one or more visual features. The second amplitude normalization method preserves the amplitude variability across the speakers, which makes it more practical in preserving the contribution of the extremas towards discrimination.

The time normalization can also be achieved in many ways. One way is to extract the visual features from the amplitude-normalized waveform as discussed in Section 3.3. The time normalization can then be applied on the extracted features by dividing the slope features by the utterance time-length, which is the number of frames between the first and last peaks as shown in Figures 3-13, 3-14, and 3-15 and presented in equation 3-1 and equation 3-2:

$$Norm\_Slope_1 = \frac{\left(\frac{y_2 - y_1}{t_2 - t_1}\right)}{(t_3 - t_1)} \quad \textbf{eq 3-1}$$

$$Norm\_Slope_2 = \frac{\left(\frac{y_3 - y_2}{t_3 - t_2}\right)}{(t_3 - t_1)} \quad \textbf{eq 3-2}$$

Where:

$Norm\_slope_1$: Normalized slope between the first and second extremas in the utterance

$Norm\_slope_2$: Normalized slope between the second and third extremas in word utterance

t_1, t_2, t_3: Time location of the 1st, 2nd, and 3rd extremas respectively.

Time normalization can also be done by mapping all utterances to the same time scale, and then extracting visual features. The second time-normalization scheme ensures that the visual features are extracted after the time-normalization is applied rather than extracting the features then applying the normalization as is the case in the first method.

One of the popular time-and-amplitude normalization techniques was introduced by Smith [63-65]. Popularity of the technique is attributed to its linearity and simplicity. The algorithm starts by applying a low pass filter with a cutoff frequency of 10 Hz to minimize the noise in the signal. The application of this low pass filter reduced the high frequency noise introduced by the pre-processing of the waveform. The amplitude normalization is achieved by dividing each word utterance by the standard deviation of the waveform. The time normalization starts with determining the starting and ending points for each word utterance and then re-sampling the word utterance on a 120 points time scale using linear interpolation.

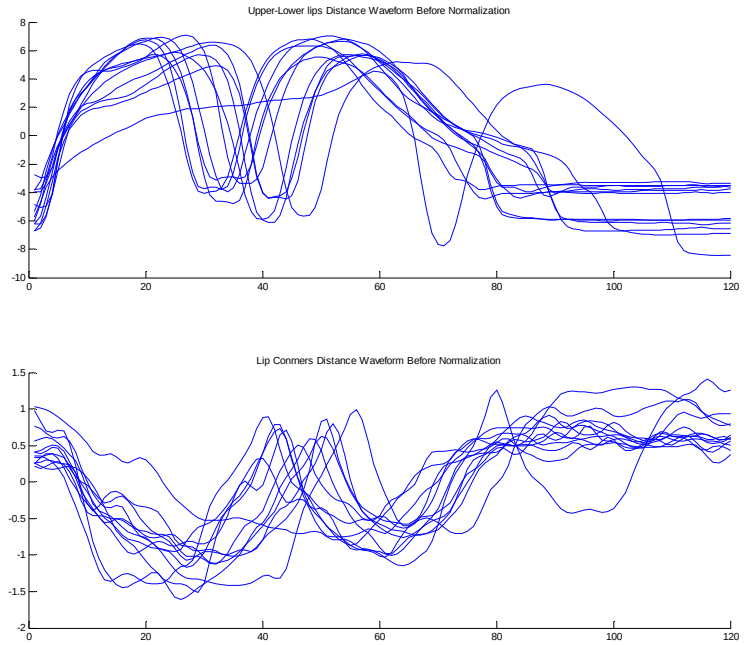

Figure 3-11 Distance waveforms associated with ten utterances of /ɑbɑ/ before normalization

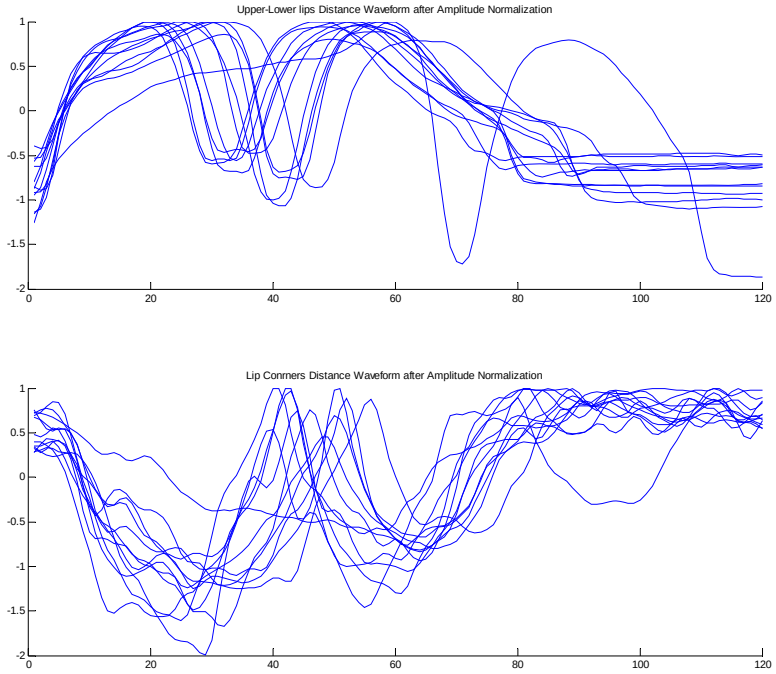

Figure 3-12 Amplitude normalization by dividing over the maximum value

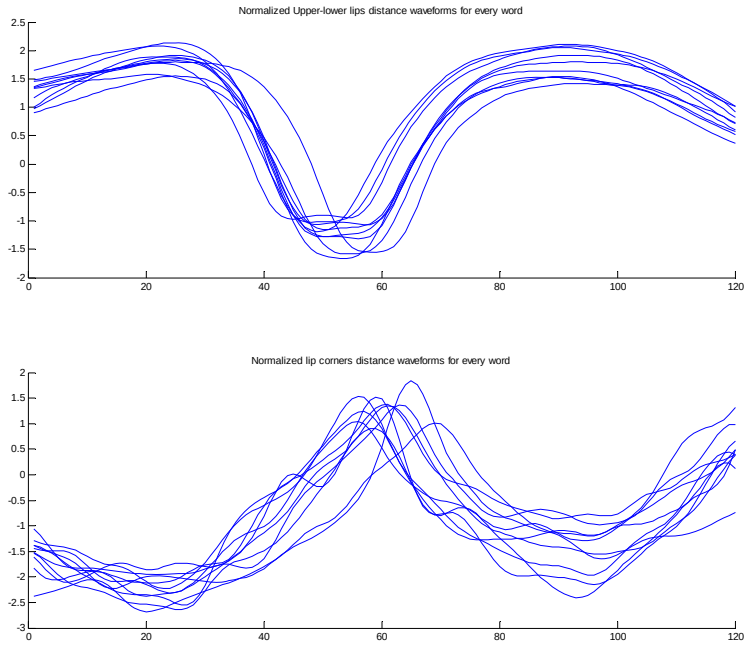

Figure 3-13 Ten word utterances after applying the Ann Smith normalization technique

Figures 3-11 displays upper-lower and lip-corners distance waveforms associated with 10 utterances coming from a speaker stating the word /ɑbɑ/. The figure shows the variations present between different utterances and affirms the need for time and amplitude normalization. Figure 3-12 displays the result of division by the maximum amplitude on these waveforms, while Figure 3-13 shows the results of applying Smith's algorithm of both time and amplitude normalizations on the waveforms.

Figure 3-12 shows how the first amplitude normalization method forces one of the extremas to be one. This may result in the redundancy of this feature in the discrimination problem.

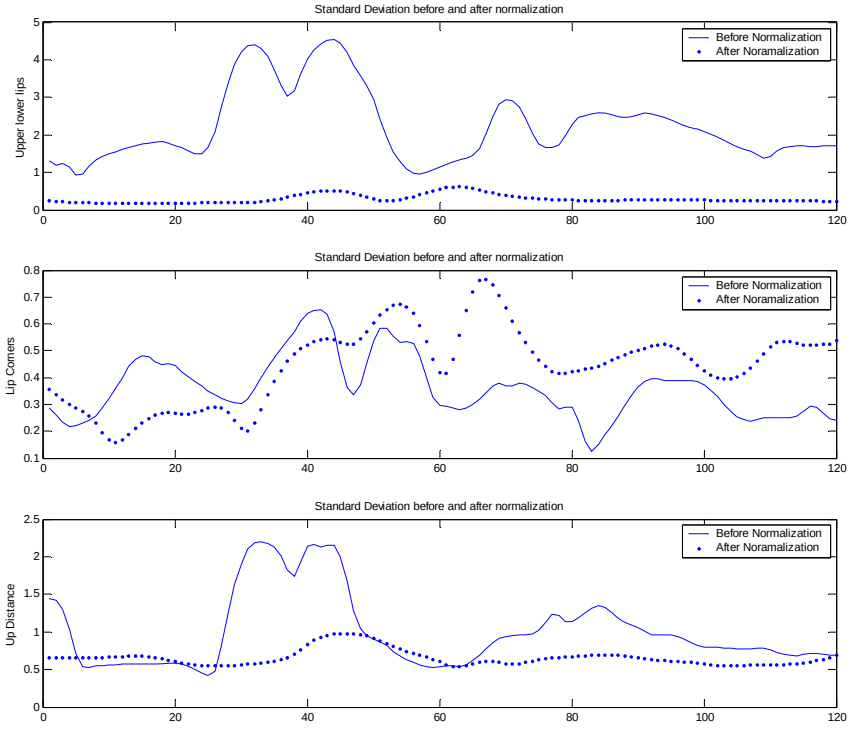

Figure 3-14 Standard deviation at each point of the 8 utterances in the 3 distance waveforms

Smith's algorithm consistently resulted in good time and amplitude normalization as shown in Figure 3-13. Figure 3-14 displays three graphs. In each graph, the standard deviation for each of the 120 samples in the waveform between the 8 different utterances for the same speaker is shown before and after normalization. The solid line represents the standard deviation across the 10 utterances before normalization while the dotted line represents the standard deviation across those points after normalization. Smith's algorithm reduced the calculated standard deviation in the upper-lower lips distance waveforms and the upper lips waveform. The algorithm is not as effective in reducing the standard deviation in lip-corners waveforms but it

does as well as the first method described. In this study, the visual features were extracted from the distance waveforms after the application of Smith's time-and-amplitude normalization.

3.3 FEATURE SELECTION AND EXTRACTION

Three normalized distance waveforms associated with each VCV word were obtained as shown in the previous section. This section discusses the process of selecting a set of parameters to capture the uniqueness in each of these waveforms. These parameters are the visual features needed for the classification problem stated in Chapter 1.

The distance waveforms capture the motion of the lips while a VCV sound is being produced. As part of preliminary work, spectral analysis and eigen space analysis to extract features were evaluated, but no patterns could be identified in either the spectrum or in the eigen vector space for the distance waveforms. Therefore, the efforts to extract features concentrated on temporal features derived from the articulator motion patterns defined in the speech production literature.

The recording device used in this study can trace different points on the face of a speaker by placing optical reflectors on them. Since the visual features are expected to represent the produced sound, the points to track around the face need to be related to the sound production. Chen's audio-visual library was based on lip-corners and lip-height as shown in Figure 1-1. Preliminary work on this library showed patterns repeating for the same word across different speakers. The literature review in Chapter 2 showed the visual features extracted from lip motion were popularly used in different applications and resulted in satisfactory results in different

areas. The articulators involved in producing the sounds selected for this study are detailed in speech production literature [12, 66, 67]. This results in the following observations:

1. The vowel /ɑ/ in the VCV sequence involves wide opening of the mouth at the start of the word, then closure of the lips to produce the consonant and then another opening of the mouth to produce the second /ɑ/ in the word.
2. The amount of closure in the mouth depends on the consonant being produced. Consonants /b/ and /v/ for example involve complete mouth closure while consonant /w/, /ð/ and /z/ involve partial closure of the mouth.
3. There is limited lip motion in producing the consonant /ð/ while there is more lip motion in producing the sound /w/.
4. The distance travelled by the upper lip in producing the consonant /v/ is greater than the distance travelled when producing the consonant /z/ [66, 67]
5. Different consonants need different speeds for the lip motion in producing them. For example, the motion speed for the lower lips while producing the sound /z/ is different from the speed of the lower lip while producing the sound /ð/ [66, 67].

The first three points are related to the distance travelled by different points around the lips when the sound is produced. This distance was represented by the maximum and minimum values detected in the upper-lower and lips corners distance waveforms. The speech production literature suggested that producing some consonants requires more motion in the upper-lip at the time of lip closure. This property was represented by the distance travelled by the upper-lip at the time of lip closure. The speed of lip motion is another characteristic that is related to the sound production. It was captured by calculating the slope between the consecutive extremas. The process of extracting the features is discussed in Sections 3.3.1 through 3.3.3.

3.3.1 Detecting Extremas in the waveforms

The extremas in the distance waveforms are related to the location of the lips during the production of the word. The distance together with the rate at which this distance changes are two parameters used to characterize these waveforms.

The extrema in upper-lower and lip-corners distance waveforms represent the maximum and minimum openings of lips during the sound production. The first extrema is related to the production of the first vowel in the VCV sequence. The second extrema is related to the production of the consonant in the VCV sequence while the third one represents the mouth opening to produce the second vowel. The amount of rounding in the mouth shape is represented by the extremas obtained from the lip corners waveform. These extremas are visually detected in each word utterance and selected using a mouse. A program stored the amplitude and the time of occurrence of the detected extremas and used these values to calculate the visual features to represent the spoken word. This process is described in the following sections.

3.3.2 Features from the upper and lower lips distance waveform

The Euclidean distance between the upper and lower lip calculated frame-by-frame is used to extract amplitude-related and time-related features. The amplitudes of the captured extremas are used as features representing the VCV sound and the slopes between these extremas are used to represent the rate at which different points around the lips move while producing the sound.

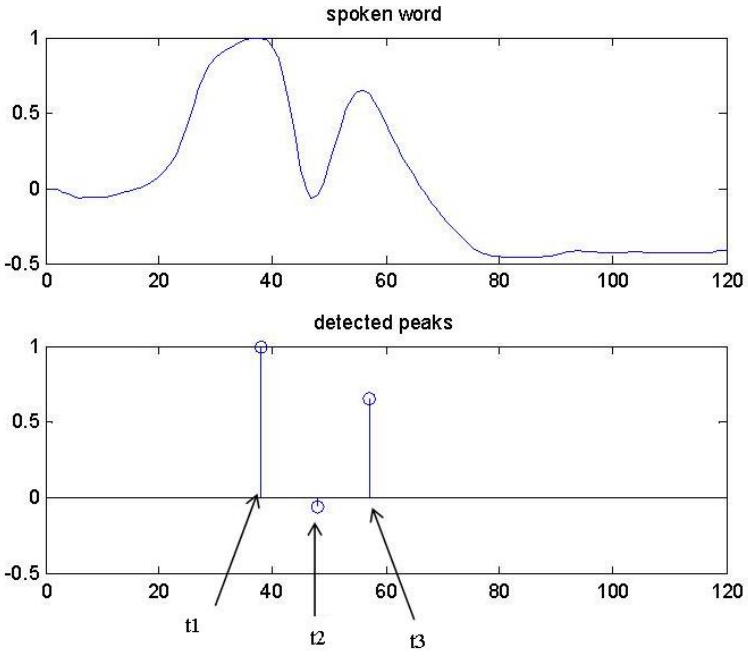

Figure 3-15 Features extracted from upper/lower distance waveform

The upper graph in Figure 3-15 shows a sample upper-lower distance waveform for the sequence /ɑbɑ/. The lower graph in the figure shows the extremas extracted from the waveform. The slope of the waveform between the extremas is calculated to capture how quickly the lips moved while stating the VCV word. The slopes are calculated by the following formulas

$$Slope_1 = \left(\frac{y_2 - y_1}{t_2 - t_1} \right) \quad \textbf{eq 3-3}$$

$$Slope_2 = \left(\frac{y_3 - y_2}{t_3 - t_2} \right) \quad \textbf{eq 3-4}$$

Where:

- y_1, y_2, y_3: Amplitude of the distance waveform at the 1st, 2nd, and 3rd extremas respectively
- t_1, t_2, t_3: Time of occurrence of the 1st, 2nd, and 3rd extremas respectively.

3.3.3 Features from the left and right lip corners distance waveform

The Euclidean distance between the lip corners waveforms calculated frame-by-frame is used to extract amplitude-related and time-related features. The amplitudes of the captured extremas are used as features representing the VCV sound and the slopes between these extremas are used to represent the rate at which different points around the lips move while producing the sound.

The upper graph in Figure 3-16 shows a sample lip corners distance waveform for the sequence /ɑbɑ/. The lower graph in the figure shows the extremas extracted from the waveform. The slope of the waveform between the extremas is calculated to capture how quickly the lips moved while stating the VCV word. The slopes are calculated by equations 3-3, and 3-4.

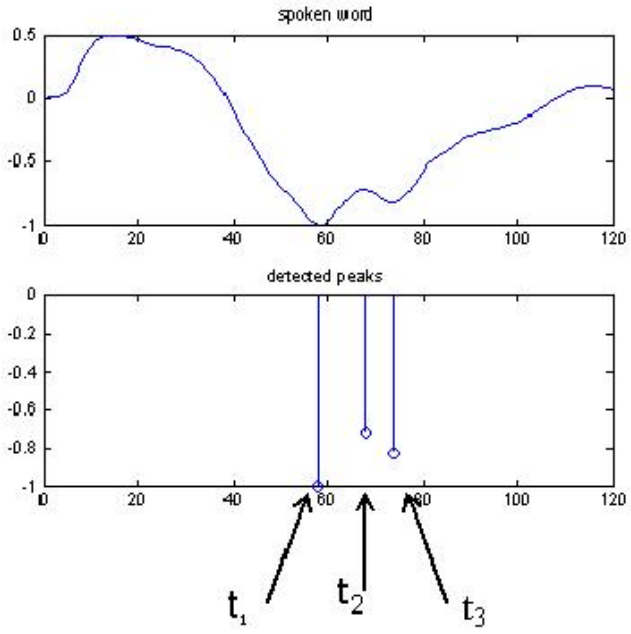

Figure 3-16 Features extracted from lip corners distance waveform

3.3.4 Features from the upper-lip distance waveform

The speech production literature suggested that producing some consonants requires more motion in the upper-lip at the time of lip closure. This property was represented by the distance travelled by the upper-lip at the time of the lip closure. Therefore, the points of interest in this waveform are chosen to match the frames at which an extrema is detected in the upper-lower lip distance waveform. Whenever an extrema is detected in the upper-lower distance waveform, the corresponding amplitude value at the upper-lip distance waveform is taken as a feature. Figure 3-17 shows an example of these features for one subject speaking the word /ɑbɑ/. The upper graph in the figure marks the extremas extracted from the upper-lower lips distance waveform. The lower graph in the figure shows the corresponding signal value extracted from the upper-lip distance waveform. In addition to the amplitude information, the slope of the waveform between the extremas is calculated to capture how quickly the lips moved while stating the VCV sequence. The slopes are calculated by Equations 3-3 and 3-4.

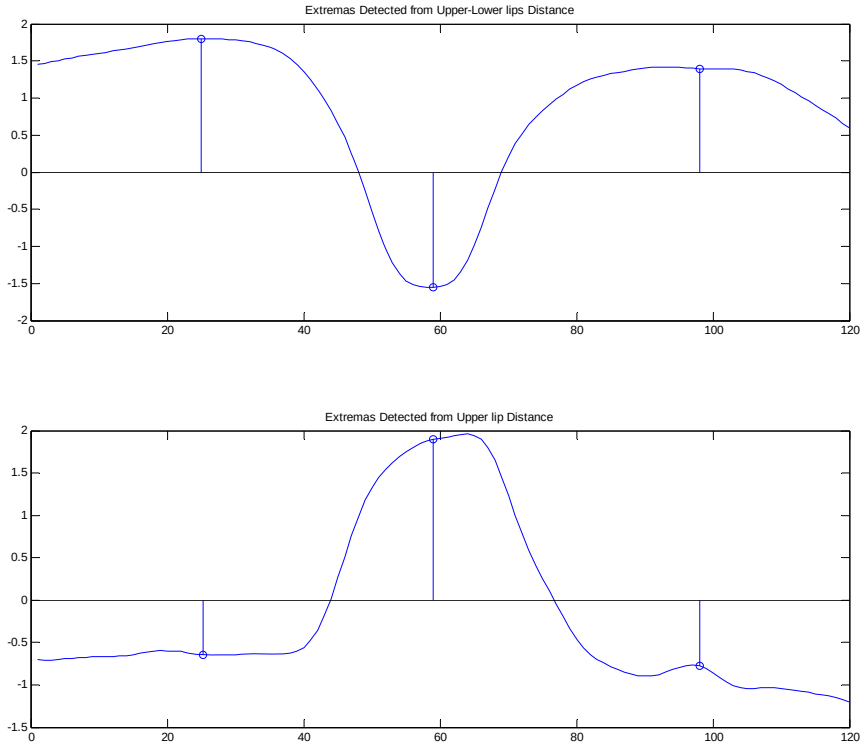

Figure 3-17 Amplitude features extracted from the upper-lips waveform

To summarize, three distance waveforms were used with every VCV word to extract the visual features. These waveforms are the upper-lower lips distance waveform, lip corners distance waveform, and waveform representing distance traveled by the upper lip in consecutive frames. Five features are extracted from each of these waveforms to make the total of 15 features for every word. The extracted features are summarized in Table 3-3:

Table 3-3 Summary of the extracted visual features

The Extracted Feature	Abbreviation
Three upper-lower lip distance extremas	UL1, UL2, UL3
Three upper lip amplitudes at locations matching the upper-lower distance extremas	UP1, UP2, UP3
Three lip corners extremas	LR1, LR2, LR3
Two features representing the slope between the consecutive extremas in the upper-lower lip distance waveform	slope_UL1, slope_UL2
Two features representing the slope between the consecutive extremas in the lip corners distance waveform.	slope_LR1, slope LR2
Two features representing the slope between the consecutive extremas in the upper lip distance waveform.	slope_UP1, slope_UP2

The effectiveness or usefulness of the above features may vary. Some features might be more important than others in the discrimination problem. The classification technique that is usually used to measure the contribution of different variables to the discrimination problem is the linear discriminant analysis, which is described in the next section.

3.4 LINEAR DISCRIMINANT ANALYSIS

Linear discriminant analysis is a statistical technique that can be used to examine whether two or more mutually exclusive groups can be distinguished from each other based on linear combinations of values of predictor variables or features (mutually exclusive means that a case can belong to only one group).

The main purpose of discriminant function analysis is to predict group membership based on a linear combination of chosen variables or features. The procedure begins with a set of observations where both group membership and the values of the features are known. The end result of the procedure is a model that allows prediction of group membership when only the features are known. A second purpose of discriminant analysis is to provide an understanding of the data set, as a careful examination of the prediction model that results from the procedure can give insight into the relationship between group membership and the features used to predict group membership. A brief discussion of Fisher's approach to discriminant analysis is discussed in the next section.

3.4.1 Discriminant Analysis Model

This section describes the development of Fisher's discriminant analysis. The material is based on lecture notes by Gutierrez-Osuna [68] and Huberty's book Applied Discriminant Analysis [69]. The concept of Fisher's discriminant functions is that given a set of independent variables or features, the analysis attempts to find linear combinations of those features that best separate the groups of cases. The set of cases separated from others are considered to be a group. The combinations of the features are called discriminant functions and have the form.

$$d_{ik} = b_{ok} + b_{1k}x_{i1} + ... + b_{pk}x_{ip} \qquad \textbf{eq 3-5}$$

where:

- d_{ik}: is the value of the k_{th} discriminant function for the i_{th} case
- p: is the number of features
- b_{jk}: is the value of the j_{th} coefficient of the k_{th} function
- x_{ij}: is the value of the i_{th} case of the j_{th} predictor
- The number of functions is equal to min(#groups-1, #features).

The procedure automatically chooses a first function that will separate the groups as much as possible. It then chooses a second function that is both uncorrelated with the first function and provides as much further separation as possible. The procedure continues adding functions in this way until reaching the maximum number of functions as determined by the number of predictors and categories in the dependent variable.

The discriminant model is based on the following assumptions:

- The features are not highly correlated with each other.
- The mean and variance of a given feature are not correlated.
- The correlation between two features is constant across groups.
- The values of each feature have a normal distribution.
- The variance-covariance matrices of the features across the various groups are the same in the population, i.e., homogeneous

3.4.2 Linear Discriminant Analysis for Two Groups

We start with a number of samples N_1 and N_2 from two independent random samples of classes w^1 and w^2 with each observation x^1, x^2 having p-dimensions with means u_1, u_2 and a

common covariance matrix ∑. The objective of the analysis is to find a scalar function "y" by projecting the samples of X^i onto a line in a way that maximizes the separability of the samples.

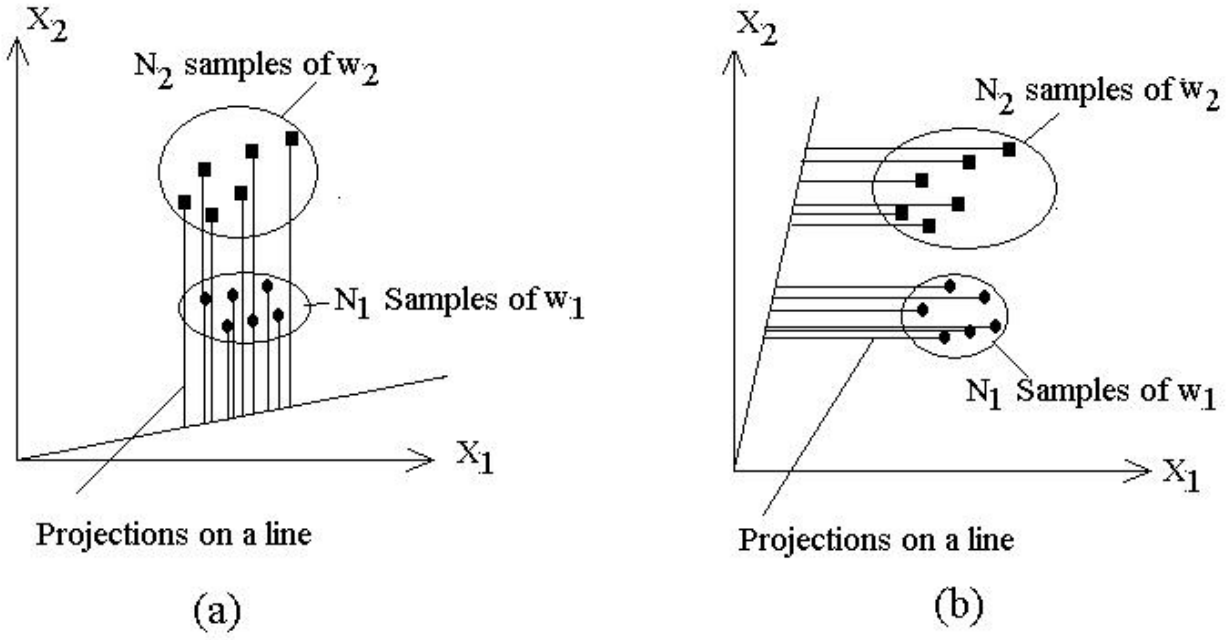

Figure 3-18 Projection of data on a line (a) Poor separability (b) Good separability

The projection of every observation is given by

$$y = w^t x \qquad \textbf{eq 3-6}$$

where "*w*" is a vector containing the coefficients for the discriminant function. Figure 3-18 shows an example for projecting a set of data belonging to two different classes on two different lines for the purpose of discriminating between both classes. Part (a) of the figure shows the result of projecting the data onto a line without achieving good separability between both classes. Part (b) of the figure shows a projection that resulted in much better separability on the line.

To explain this concept further, a separability measure needs to be defined. In order to achieve this, the mean vector for each class in x and y feature space is defined as

$$\mu_i = \frac{1}{N_i} \sum_{x \in w_i} x \qquad \textbf{eq 3-7}$$

This means that the mean along the line of projection is given by

$$\hat{\mu}_i = \frac{1}{N_i} \sum_{y \in w_i} y = \frac{1}{N_i} \sum_{y \in w_i} w \; x = w^t x \qquad \textbf{eq 3-8}$$

Now we could chose the distance between the projected means as our objective function

$$J(w) = | \hat{\mu}_1 - \hat{\mu}_2 | = | w^t (\mu_1 - \mu_2) | \qquad \textbf{eq 3-9}$$

The distance between the projected means is not a very good measure because it does not take the standard deviation within the classes into account. Fisher presented a solution to this problem by suggesting maximizing a function that represents the difference between the means, normalized by a measure of the within-class scatter. Fisher defined the scatter for each class as

$$s_i^2 = \sum_{y \in w_i} (y - \hat{\mu}_i)^2 \qquad \textbf{eq 3-10}$$

Then he defined the within-class scatter of the projected samples to be

$$within - class - scatter = s_1^2 + s_2^2 \qquad \textbf{eq 3-11}$$

The Fisher linear discriminant is defined as the linear function $\mathbf{w^t x}$ that maximizes the criterion function

$$J(w) = \frac{| \hat{\mu}_1 - \hat{\mu}_2 |^2}{s_1^2 + s_2^2} \qquad \textbf{eq 3-12}$$

Therefore, we would be looking for a projection where examples from the same class are projected very close to each other and at the same time, the projected means of different classes are as far apart as possible. Fisher's solution to the above problem is given by

$$w^* = S_w^{-1}(\mu_1 - \mu_2) \qquad \textbf{eq 3-13}$$

and the within class scatter matrix S_w is given by

$$S_w = S_1 + S_2 \qquad \textbf{eq 3-14}$$

Equation 2-9 is usually known as the Fisher Linear Discriminant function, which represents a specific choice of direction for the projection of the data down to one line. This equation can be generalized for C-class problems, and the next section will discuss this process.

3.4.3 Linear Discriminant Analysis, C-classes

The solution presented by Fisher shown in equation 2-9 can be extended for a general situation that involves C-classes. In this case, we will seek (C-1) projections [y_1,y_2,y_3, ….. y_{c-1}] by means of (C-1) projection vectors w_i which can be arranged by columns into a projection matrix W=[w_1|w_2|w_3|…|w_{c-1}] so the problem becomes

$$y_i = w_i^t x \Rightarrow y = W^t x \qquad \textbf{eq 3-15}$$

The solution to the above problem is given by the following equation

$$W^* = [w_1^* \mid w_2^* \mid ..... \mid w_{c-1}^*] \Rightarrow (S_B - \lambda_i S_w) w_i^* = 0 \qquad \textbf{eq 3-16}$$

where S_B is defined as the generalized between-class-scatter and given by:

$$S_B = \sum_{i=1}^{c} N_i (\mu_i - \mu)(\mu_i - \mu)^t \qquad \textbf{eq 3-17}$$

with

$$\mu = \frac{1}{N} \sum_{\forall x} x = \frac{1}{N} \sum_{x \in w_i} N_i \mu_i \qquad \textbf{eq 3-18}$$

S_w, which is the generalization of the within-class scatter, is

$$S_w = \sum_{i=1}^{C} S_i \qquad \textbf{eq 3-19}$$

with

$$S_i = \sum_{x \in w_i} (x - \mu_i)(x - \mu_i)^t \qquad \textbf{eq 3-20}$$

and

$$\mu_i = \frac{1}{N_i} \sum_{x \in w_i} x \qquad \textbf{eq 3-21}$$

Equation 3-16 simply means that the projections with maximum class separation are the eigen vectors corresponding to the largest eigen values of $S_w^{-1} S_B$.

3.4.4 Stepwise Discriminant Analysis

Some variables may contribute more to the discrimination problem than others. Variables with very little contribution to the discrimination problem can be discarded to reduce the complexity and dimensions of the problem. The stepwise discriminant analysis is a technique used to test which variables contribute more to the discrimination function. It can help in reducing the dimensions of the problem by discarding variables that have insignificant contribution to the discrimination function.

Before the stepwise process, a statistical measure for evaluating each variable in the analysis, together with a significance level of F values that a variable must have to enter a model or be removed from the model, must be developed. Once the criteria and the F values are chosen, the Linear Discriminant Function (LDF) is estimated for all variables. The process proceeds as follows [70]:

- The variable that best meets the criteria is entered into the analysis.
- The remaining variables are tested again and the variable with the best value for the selection criteria is added to the analysis.
- The variables in the model are tested to check if any meet the removal criteria and variables meeting the criteria are removed.
- The process of evaluating variables not in the model is repeated until all variables have been tested for entry or removal.
- The process is terminated when no more variables meet the entry or removal criteria.

In this work, the statistical measure chosen is Wilk's lambda which is the ratio of the generalized within-class-scatter given by equation 3-19 to the generalized over all scatter given by equation 3-17. The change in Wilk's lambda for a model if a variable is added or removed is calculated by the following formula

$$F_{change} = \left(\frac{n-g-p}{g-1} \right) \left(\frac{1 - \frac{\lambda_{p+1}}{\lambda_p}}{\frac{\lambda_{p+1}}{\lambda_p}} \right) \qquad \textbf{eq 3-22}$$

Where:

- λ_p is the Wilk's lambda value before adding the new variable
- λ_{p+1} is the Wilk's lambda value with the added variable
- g is the number of groups
- p is the number of independent variables entered in the stepwise analysis
- n total sample size

There are many software packages available to perform the discriminant analysis. One of those packages is SPSS version 16 which is a statistical software package that can perform the

above calculations and provide useful statistics to better understand the strength of the discrimination and the distribution of the data.

4.0 RESULTS

Experiments were conducted to evaluate the ability of visual cues to classify visemes associated with VCV words. The first section of this chapter presents the results of developing the linear discriminant functions (LDA) needed for the classification (training phase) with all speakers involved in the training phase. The second section presents the results of testing the developed functions. The third section shows the step-wise linear discriminant analysis results, and the final section presents the results of training and testing discrimination functions that were built for each individual speaker.

4.1 TRAINING AND TESTING THE CLASSIFIER

There are two methods for generating the training data by partitioning the VCV sequences coming from the 18 speakers shown in bold fonts in Table 3-1. These two methods are speaker-based training and word-based training. In speaker-based training, VCV words from 9 speakers are used to develop the LDA functions and VCV words from the remaining 9 speakers are used to test the resulting functions. In word-based training, all the available VCV words are divided into two equal parts. One is used for developing the LDA functions and the other one is used for testing them.

4.1.1 Speaker Based Training

LDA functions were developed and tested for different numbers of discrimination classes using speaker-based data analysis.

4.1.1.1 Speaker-Based Training with 12 Classes

In this part, each VCV word was treated as a separate class and the SPSS linear classification algorithm was applied to calculate the LDA functions needed to discriminate between those 12 classes.

A prerequisite for SPSS analysis is to ensure the validity of the assumptions listed in Section 3.4.1. This can be achieved by many tests performed by the SPSS package. The Wilk's lambda test seeks to confirm the assumption of un-equal means between the LDA functions. It tests the null hypothesis that the population means for all the discrimination functions are equal in all the classes. If the hypothesis is accepted, then the discrimination functions represent nothing more than the sampling variability. SPSS calculates the value of lambda for different functions. If the significance level for the function is small, then the null hypothesis is rejected. The first step of the test calculates Wilk's lambda for all 11 functions according to equation 4-1

$$\lambda = \prod_{\forall i} \frac{S_w}{S_B} \qquad \text{eq 4-1}$$

where:

S_w is within-class scatter and given by eq 3-19

S_B is overall-class-scatter and given by eq 3-17

i is the i^{th} discrimination function, i=1:11

In the following steps, the test excludes one function at a time and calculates the Wilk's lambda for the remaining functions. Table 4-1 shows the results of this test. The small significance values shown in the 3rd column of the Table indicates that the null hypothesis is rejected for the first eight functions. The differences between the means in the remaining three functions are not sufficient, indicating that these functions have small contribution towards the discrimination.

Table 4-1 Testing equality of means for speaker based training analysis

Test of Function(s)	Wilks' Lambda	Sig.
1 through 11	.011	.000
2 through 11	.059	.000
3 through 11	.189	.000
4 through 11	.309	.000
5 through 11	.486	.000
6 through 11	.690	.000
7 through 11	.838	.000
8 through 11	.932	.003
9 through 11	.968	.163
10 through 11	.984	.310
11	.994	.396

Another assumption involved in the LDA analysis is that samples represent a multivariate normal distribution with equal covariance matrices in the population. SPSS performs the Box M test to verify the validity of this assumption. Box M tests the null hypothesis that the covariance matrices for the features are equal. The SPSS literature states that for sample sizes of more than

40, the normality test may detect statistically significant but unimportant deviations from normality [70].

Table 4-2 Test of equality of covariance matrix between groups

Class	Rank	Log Determinant
/ɑbɑ/	15	-70.745
/ɑvɑ/	15	-65.853
/ɑðɑ/	15	-56.453
/ɑwɑ/	15	-63.055
/ɑdɑ/	15	-64.485
/ɑzɑ/	15	-60.873
/ibi/	15	-50.944
/ivi/	15	-63.958
/iði/	15	-67.769
/iwi/	15	-57.317
/idi/	15	-64.722
/izi/	15	-53.074
Pooled within-groups	15	-51.497

Table 4-2 shows the results for the test of equality of covariance matrices between groups. The second column of Table 4-2 shows that the covariance matrices for each of the 12 classes are full ranked. However, the 3rd column of Table 4-2 shows that the log determinant for the covariance matrix associated with every class is not always close to the overall covariance matrix. The significance results for the Box M test presented in Table 4-3 shows that the null

hypothesis is rejected. The literature associated with the SPSS software stated that small variability in the data available in a large sample can result in failing the normality test, but the LDA analysis can still be used to discriminate between the classes.

Table 4-3 Tests null hypothesis of equal population covariance matrices.

Box's M		8611.260
F	Approx.	5.996
	df1	1320.000
	df2	547550.691
	Sig.	.000

Table 4-4 shows how well the training process does in classifying all 12 classes. The average percent of correct classification is 55.3%, which is much higher than chance (8%). In addition, the features discriminated the following classes (/ɑbɑ/, /ɑwɑ/, /eðe/, /iwi/, /ɑðɑ/, /izi/, and /iði/) with a correct percent of classification above the average performance. The Table shows that sounds are confused with each other at different rates.

Two sequences VCV_1 and VCV_2 are mutually confused if utterances of VCV_1 are mis-classified as VCV_2 and utterances of VCV_2 are mis-classified as VCV_1. For example, VCV sequences for the same consonant and different vowels are often mutually confused as shown in cells with thick boarder lines. Sixteen of the 29 mis-classified sequences of /ɑvɑ/ were classified as /ivi/. In addition, 8 of the 34 mis-classified sequences of /ivi/ were classified as /ɑvɑ/. This mutual confusion is also present between other sequences such as /ɑðɑ/ and /iði/. This mutual

confusion is consistent with the results of common viseme-to-phoneme mappings presented in Table 2-11.

The Table shows other mutual confusions between sounds associated with consonant pairs /d/, and /z/, in addition to /ɑbɑ/ and /ɑvɑ/. The viseme-to-phoneme mappings in Table 2-11 were obtained based on visual observations of VCV sequences. Sequences that had similar visible articulators were confused with each other and assigned to the same visible class by different observers. The visual features used in this study were extracted from the motion of the visible articulators. The confusion results in Table 4-4 show that the classifier has confusion patterns similar to the ones shown by studies involving human identification of visemes. These confusions will be discussed further as the effect of merging mutually confused classes together on the performance of the classifier is studied.

SPSS performs additional analyses that help in understanding the discrimination problem. Table 4-5 presents the contribution of each of the resulting 11 functions in discriminating between the variables. The first column shows the function number, the 2nd column shows the percentage of explained variance by the function across the data, and the 3rd column presents the cumulative explained variance achieved by adding the scores for functions above that row. Results of this test indicate that 97% of the variance in the data can be explained by the first six functions.

There are two more important results generated by the SPSS package. The first one is called the structural matrix, which shows the contribution of each visual feature towards the discrimination problem. The second important result is the coefficients of the discrimination functions. Tables 4-6 and 4-7 show these results respectively.

Table 4-4 Classification results and the confusion matrix 12-class speaker based

Class		Predicted Group Membership												
		/aba/	/ava/	/aða/	/awa/	/ada/	/aza/	/ibi/	/ivi/	/iði/	/iwi/	/idi/	/izi/	Total
Original Count	/aba/	**66**	5	0	0	0	0	1	0	0	0	0	0	72
	/ava/	11	**43**	0	0	1	0	1	16	0	0	0	0	72
	/aða/	3	4	**20**	0	6	6	1	8	10	1	10	3	72
	/awa/	12	0	0	**56**	0	0	0	1	0	2	0	1	72
	/ada/	1	5	16	0	**17**	12	0	5	3	0	6	7	72
	/aza/	4	6	6	1	8	**24**	0	4	10	0	8	1	72
	/ibi/	13	0	2	1	4	0	**49**	1	1	0	0	1	72
	/ivi/	4	8	0	0	4	5	0	**38**	8	0	0	5	72
	/iði/	0	0	10	0	1	2	0	0	**42**	1	16	0	72
	/iwi/	3	1	4	5	0	1	0	3	1	**54**	0	0	72
	/idi/	1	0	4	0	7	8	0	3	15	1	**33**	0	72
	/izi/	0	4	9	1	4	3	0	4	2	0	9	**36**	72
	%	91.7	59.7	27.8	77.8	23.6	33.3	68.1	52.8	58.3	75.0	45.8	50.0	

The columns of Table 4-6 represent the discrimination functions, and the rows represent the correlation between the feature and the score of a discriminating function. The higher the correlation is, the more contribution this feature has towards the discrimination score provided by the function. Features with the highest contribution in each function are picked up by SPSS program and shown in bold fonts.

The second extrema in the lip-corners distance is the most significant feature in the discrimination score provided by the 1st discriminant function. The order in which features are arranged for the 1st function is not the same for the remaining functions, which means that a

certain feature may contribute more in the discrimination by a given function while playing little role in discriminating in other functions.

Table 4-5 Contribution of the discriminant functions towards the classification problem

Function	% of Variance	Cumulative %
1	49.4	49.4
2	21.8	71.2
3	11.9	83.1
4	6.3	89.4
5	5.1	94.5
6	2.6	97.1
7	1.6	98.7
8	.5	99.2
9	.3	99.6
10	.3	99.8
11	.2	100.0

Table 4-5 showed that for a 97% overall accuracy with the training data, it would be enough to consider the first 6 functions and discard the remaining ones. This suggests that the following features (Slope_LR2, Slope_LR1, LR2, Slope_UP2, UP2, Slope_UP1, Slope_UL2, UL1, UL3, UL2, LR3) contribute more to the discrimination than the remaining ones (slope_UL1, UP1, UP3, and LR1).

Table 4-6 The contribution of each feature towards the discrimination (Structural matrix)

Feature	Function										
	1	2	3	4	5	6	7	8	9	10	11
LR2	**-.563***	**-.522**	.190	**.401**	-.158	.034	.087	-.023	-.213	.096	.149
Slope_UP2	**-.529***	.177	**-.445**	-.163	.022	.096	-.278	-.286	.019	.228	-.085
Slope_UP1	**.528***	-.351	.375	.050	-.116	.127	.081	-.133	.113	-.330	.045
UP2	**.501***	-.202	.306	-.121	-.290	.265	.346	.071	-.287	.250	-.306
Slope_LR2	**.471**	**.690***	-.136	.296	.028	.313	-.025	-.042	-.074	-.137	-.060
Slope_LR1	**-.473**	**-.669***	.126	-.146	-.008	-.014	.032	.168	-.243	-.061	.046
LR1	.168	**.460***	-.029	.211	.013	-.269	.119	-.028	-.209	.151	.296
UL3	-.020	-.075	-.172	-.074	**.639***	**.411**	-.178	.262	-.054	-.175	.342
UL1	-.095	-.091	-.169	-.045	**.601***	**.494**	-.252	.020	.048	-.348	-.063
UL2	-.250	.291	.040	-.350	**.409**	**.465***	.056	.216	.087	-.369	.266
Slope_UL1	-.180	.366	-.109	-.261	-.130	.109	**.475***	.132	.241	-.147	.340
Slope_UL2	.284	-.263	.241	.352	.358	-.243	**-.429***	.080	.015	.307	.049
LR3	.042	.326	.147	**.411**	.161	.189	.190	**-.606***	-.265	-.001	.237
UP1	-.024	.124	.055	-.064	.019	.092	**.417**	.202	-.255	**.687***	**-.404**
UP3	.029	.042	.147	-.167	-.101	.191	.078	-.133	-.338	.356	**-.514***

Table 4-7 shows the coefficients of the 11 discrimination functions calculated using the training set. The performance of the discriminator is tested by obtaining the dot product between an unknown record and the coefficients of all the functions and then assigning that record to the class associated with the function that resulted in the highest score [70].

Table 4-7 Classification function coefficients

features	Function											
	1	2	3	4	5	6	7	8	9	10	11	12
UL1	1.211	2.483	.560	2.500	.858	.236	-.609	1.831	-.560	-.110	.565	2.436
UL2	-5.624	-8.215	-1.015	-7.009	-1.785	-2.039	-2.108	-6.111	.652	-3.233	-.438	-3.672
UL3	6.229	7.341	2.892	7.034	3.244	3.788	4.983	6.193	1.600	4.675	1.797	5.249
LR1	-4.034	-2.710	-2.709	-1.492	-3.211	-2.910	-3.925	-2.135	-2.256	-1.469	-3.331	-2.922
LR2	2.044	2.217	1.776	-3.724	2.812	2.511	1.309	1.541	1.743	.319	3.285	2.480
LR3	-2.414	-2.680	-2.548	-1.055	-2.690	-2.149	-.966	-2.103	-1.813	-2.465	-1.875	-1.753
UP1	-.512	.935	.189	-.198	.757	-.656	1.153	1.611	.120	.106	.558	2.054
UP2	4.111	-1.888	.007	3.157	.077	.259	-.662	-1.900	.551	1.046	-.339	-2.021
UP3	-3.758	-.566	-.671	-2.863	-1.708	-.648	-.358	-.336	-1.221	-1.459	-.793	-.416
Slope_UL1	5.891	34.763	-22.292	23.503	-14.189	-34.393	-36.739	19.305	-26.315	-9.467	-18.495	17.989
Slope_UL2	17.023	-37.834	29.650	22.780	54.386	29.698	72.226	-22.957	42.830	52.321	32.218	12.492
Slope_LR1	-14.972	-13.460	-5.054	12.706	-20.396	-14.944	-22.420	-7.375	-13.218	-29.758	-29.647	-7.793
Slope_LR2	17.473	14.018	6.296	-13.222	16.678	12.961	-9.184	9.621	6.737	68.177	15.272	21.009
Slope_UP1	14.955	31.766	-6.844	20.020	13.488	-24.260	74.064	27.627	-11.702	18.793	6.345	38.253
Slope_UP2	48.609	-29.999	-6.792	27.674	28.107	7.123	-48.588	-27.771	9.683	-7.070	7.550	-.204
(Constant)	-23.418	-19.112	-12.345	-23.949	-14.781	-13.717	-23.829	-13.721	-7.625	-18.136	-9.602	-16.592

4.1.1.2 Speaker-Based Training with 6 Classes

Owen's work summarized in Table 2-9 showed that the visual identification of the VCV sounds was not affected by the change in the vowel involved in the VCV sequence. This means that viewers in Owen's experiments assigned both /ɑbɑ/ and /ibi/ to the same viseme class. Results of the 12 classes shown in Table 4-4 indicate that when the vowels were considered different classes, the classifier had an overall 55.3% of correct classification. However, the results showed a level of mutual confusion between consonants associated with both vowels. In this section, the speaker-based training is performed assuming that sounds coming from different

vowels with the same consonant belong to the same class. This reduced the number of classes to the six shown in Table 4-8:

Table 4-8 Six classes resulting from combining words for the same vowel

Class #	VCV Sounds
1	/ɑbɑ,ibi/
2	/ɑvɑ,ivi/
3	/ɑðɑ,iði/
4	/ɑwɑ,iwi/
5	/ɑdɑ,idi/
6	/ɑzɑ,izi/

The classification results for this training set are shown in Table 4-9 and the average percentage of correct classification was 65.2%.

Merging the vowels into one class increased the performance in recognizing some of the sounds. For example, the recognition rates for /ɑvɑ/, /ɑðɑ/ in the 12 class case were 59.7% and 27.8% respectively. These percentages increased to 74.3% and 68.1% respectively when the vowels were merged together. Classes associated with the consonants /b/ and /w/ had the highest recognition score in both 12 class and 6 class cases. Classes associated with consonants /d/ and /z/ had the poorest performance in both 6 and 12 class configurations. In the 6 class configuration, 33 utterances involving the consonant /d/ were classified as /z/, and 22 utterances involving the consonant /z/ were classified as /d/. The total number of mis-classified utterances for both sounds was 173, and 55 of them (32%) were mutually confused. This confusion is

shown in cells marked with thick boarders in Table 4-9. The mutual confusion between visemes representing sequences /z/ and /d/ is a result of them having similar visual articulators as discussed in the previous section.

Table 4-9 Classification results and the confusion matrix 6 class speaker based

Classes		Predicted Group Membership						
		/ɑbɑ,ibi/	/ɑvɑ,ivi/	/ɑðɑ,iði/	/ɑwɑ,iwi/	/ɑdɑ,idi/	/ɑzɑ,izi/	Total
Original Class	/ɑbɑ,ibi/	**130**	5	4	1	1	3	144
	/ɑvɑ,ivi/	13	**107**	10	1	10	3	144
	/ɑðɑ,iði/	4	10	**98**	2	9	21	144
	/ɑwɑ,iwi/	19	8	4	**113**	0	0	144
	/ɑdɑ,idi/	4	23	28	1	**55**	33	144
	/ɑzɑ,izi/	2	17	41	2	22	**60**	144
	%	90.3	74.3	68.1	78.5	38.2	41.7	

There were other confusion patterns appearing in the table between classes involving the consonant /ð/ and the pair /d,z/. 99 of the 224 mis-classified utterances involving these sounds were mutually confused. In addition, despite that /b/ and /v/ had recognition rates higher than the overall performance, 18 of the 51 mis-classified utterances for /b/ and /v/ were mutually confused.

The structural matrix associated with the resulting classification functions is shown in Table 4-10. Features that have high correlation with the score of each function are shown in bold

fonts. The top six significant variables for the first function are (LR2, Slope_UP1, Slope_UP2, UP2, Slope_LR1, and Slope_LR2). In the second function, Slope_UL1 became important. Slope_UP2, and UP2 did not have high correlation with the score associated with the function. The results in Tables 4-9 and 4-10 are discussed further in Chapter 5.

Table 4-10 The contribution of each feature towards the discrimination (Structural matrix)

features	Function				
	1	2	3	4	5
Slope_UP2	**-.506***	-.392	-.213	.197	.134
UP2	**.502***	.374	.346	-.150	-.050
LR2	**-.569**	**.683***	-.171	-.003	-.025
Slope_LR1	**-.430**	**.577***	-.012	-.098	**-.475**
Slope_LR2	**.412**	**-.524***	-.007	.291	.306
Slope_UP1	**.513**	**.520***	.227	.018	-.003
Slope_UL1	-.163	**-.402***	.196	-.312	-.002
LR1	.158	-.371*	-.072	-.111	.103
Slope_UL2	.252	.337*	-.161	.297	.011
UP3	.032	.011	.246*	.065	.042
UL1	-.090	-.049	-.141	**.673***	-.352
UL3	-.020	-.069	-.142	**.568***	**-.555**
LR3	.041	-.134	-.044	.354*	.218
UP1	-.019	-.074	.092	-.146*	.015
UL2	-.235	-.368	.317	.358	**-.401***

4.1.1.3 Speaker-Based Training with 5 Classes

One of the objectives of this study was to evaluate how well these features could discriminate between the /d/ and the /z/ sounds. Both of these sounds were associated with the same class shown in Table 2-11. Table 4-9 shows that these two sounds are highly confused with each other. Utterances of these sounds are merged together resulting in the 5-class situation shown in Table 4-11. The classification results for this training set are shown in Table 4-12. The average percentage of correct classification was 72.2%

Table 4-11 Five classes resulting from combining /VdV/ with /VzV/

Class #	VCV Sounds
1	/ɑbɑ,ibi/
2	/ɑvɑ,ivi/
3	/ɑðɑ,iði/
4	/ɑwɑ,iwi/
5	/ɑdɑ,idi,ɑzɑ,izi/

Table 4-12 Classification results and the confusion matrix 5 class speaker based

Classes		Predicted Group Membership					
		/ɑbɑ,ibi/	/ɑvɑ,ivi/	/ɑðɑ,iði/	/ɑwɑ,iwi/	/VzV,VdV/	Total
Original Class	/ɑbɑ,ibi/	**128**	4	3	1	8	144
	/ɑvɑ,ivi/	13	**98**	8	1	24	144
	/ɑðɑ,iði/	2	5	**97**	3	37	144
	/ɑwɑ,iwi/	19	6	3	**112**	4	144
	/VzV,VdV/	6	35	58	1	**188**	288
	%	88.9	68.1	67.4	77.8	65.3	

The performance of individual classes between the 6-class and 5-class configuration did not change except for the class involving consonants /d/ and /z/. These two consonants had 38% and 42% correct perception in the 6 class case and merging them improved the overall performance to 65.3%. Classes associated with the consonants /b/ and /w/ continued to have the highest correct percentage score. The mutual confusion observed in the 6 class configuration between classes associated with consonants /d/, /z/ and /ð/ remained (69% of the mis-classified utterances were mutually confused).

Table 4-13 The contribution of features towards the discrimination (Structural matrix)

Feature	Function			
	1	2	3	4
Slope_UP2	**-.491***	**-.410**	-.234	.244
UP2	**.486***	.386	.358	-.224
LR2	**-.600**	**.644***	-.149	-.053
Slope_UP1	**.497**	**.540***	.244	-.043
Slope_LR1	**-.465**	**.538***	.045	-.133
S ope_LR2	**.446**	**-.477***	-.045	.242
Slope_UL1	-.134	**-.411***	.196	-.337
UL2	-.214	-.367*	.347	.298
LR1	.179	-.354*	-.092	-.134
UP3	.045	.015	.244*	.014
UL1	-.098	-.040	-.115	**.603***
UL3	-.038	-.061	-.106	**.527***
Slope_UL2	.215	.348	-.169	.390*
LR3	.072	-.111	-.054	.199*
UP1	-.009	-.074	.092	-.159*

The structural matrix associated with the resulting discrimination functions is shown in Table 4-13. There was no change in the order of importance for the features in the first function for both 5-class and 6-class cases. Slope_UP2 became a significant feature in the second function and LR2 remained the feature with the highest correlation with the discrimination score in the first function for all 3 class combinations presented. These results are discussed further in Chapter 5.

4.1.1.4 Speaker-Based Training with 3 Classes

Table 4-12 shows that the three classes /aða,eðe/, and /Vz,dV/ have a great deal of mutual confusion. The confusion between classes /aba,ebe/ and /ava,eve/ continued to appear in all class configurations tested. In this section, these classes are combined and a discriminator is designed based on the resulting training data. Table 4-14 shows the resulting classes after combining the classes.

Table 4-14 Three classes resulting from combining /VdV/ with /VzV/

Class #	VCV Sounds
1	/ɑbɑ,ibi/, /ɑvɑ,ivi/
2	/ɑðɑ,iði/, /ɑdɑ,idi/, /ɑzɑ,izi/
3	/ɑwɑ/,/iwi/

The classification results for this training set is shown in Table 4-15 The average percentage of correct classification was 84.4%

Classes associated with consonants /b/ and /w/ continued to have the highest classification score. In addition, in the 3 classes configuration confusion between consonants involving partial mouth closure like /w/ had confusion with classes involving complete mouth contact like /b/.

Table 4-15 Classification results and the confusion matrix 3 class speaker based

Classes		Predicted Group Membership			
		/ɑb,vɑ,ib,vi/	/ɑwɑ/,/iwi/	/ɑð, d, zɑ/ /ið,d,z,i/	Total
Original Class	/ɑb,vɑ,ib,vi/	**240**	1	47	288
	/ɑwɑ/,/iwi/	28	**110**	6	144
	/ɑð, d, zɑ/ /ið,d,z,i/	47	6	**379**	432
	%	83.3	76.4	87.7	

The structural matrix associated with the discrimination functions is shown in Table 4-16. The structural matrix shows that LR2, Slope_LR1, and Slope_LR2 have the highest correlation with the score of the first and second functions. The second function's score is also correlated with UL2, Slope_UP1, and Slope UL1. The results associated with the speaker training are discussed further in next chapter.

Table 4-16 The contribution of each features towards the discrimination (Structural matrix)

Features	Function	
	1	2
LR2	**.666***	**-.603**
Slope_LR1	**.529***	**-.498**
Slope_LR2	**-.502***	**.474**
Slope_UP2	.399*	.364
UP2	-.377*	-.288
UL2	.227	**.557***
Slope_UP1	-.371	**-.415***
Slope_UL1	.122	**.415***
Slope_UL2	-.216	-.361*
LR1	-.223	.277*
LR3	-.061	.164*
UL1	.084	.109*
UL3	.009	.095*
UP3	-.011	.087*
UP1	.015	.085*

4.1.2 Testing Models developed by Speaker-based Training

Discriminant models developed using the training set were tested with features extracted from word utterances coming from 9 speakers who were different from the speakers used to develop the discriminant models. The results of testing are shown in Table 4-17, where the VCV sequences included in each of the classes are given in Tables 4-2, 4-8, 4-11, and 4-14 respectively.

Performance of the developed LDA functions in the testing set for the 6-class, 5-class and 3-class configurations were close to the training results. The test results for the 12-class configuration were 12% lower than those of training class. This high drop in the test set can be attributed to the difficulty present in distinguishing both vowels from each other.

Table 4-17 Testing the Fisher functions developed in speaker based training

Number of Classes	% correct Classification	
	Training	Testing
Twelve	55.3	43.1
Six	65.2	59.91
Five	72.1	69.93
Three	84.4	83.37

The confusion matrices associated with the testing phase are presented in Tables 4-18 through 4-21. Classes /ɑbɑ/, /ɑwɑ/ and /ewe/ had a high percentage of correct classification when the testing was done for 12 classes, which is consistent with training results. This indicates that the features captured the uniqueness of the visual cues associated with these sounds fairly well. The pairs of classes </ɑðɑ/, /iði/>, </ɑwɑ/, /iwi/>, </ɑdɑ/,/idi/>, </ibi/,/ɑbɑ/>, and </ivi/,/ɑvɑ/> were mutually confused. This confusion is consistent with the common viseme-to-phoneme mappings shown in Table 2-9 in which VCV sequences of the consonant were assigned to the same viseme class. The mutual confusion between VCV sequences involving /d/ and /z/ appeared in the test set as it appeared in the training set. Comparing the testing results from Table 4-18 with the training results of Table 4-4, the confusion patterns between utterances in both sets are very similar to each other.

When the number of classes was reduced to 6, the confusion patterns discussed in the 6-class training set continued to exist in the 6-class testing set. These results are shown in Table 4-19. The mutual confusion between classes /d/ and /z/ is consistent with the results of the common viseme-to-phoneme mappings shown in Table 2-11. The overall percent of correct classification in the test set was 59.91%. In addition, classes involving consonants /b/ and /v/ were mutually confused as they were in the training set.

Table 4-18 Confusion matrix for the 12 class testing phase

Class		Predicted Group Membership												
		/ɑbɑ/	/ɑvɑ/	/ɑðɑ/	/ɑwɑ/	/ɑdɑ/	/ɑzɑ/	/ibi/	/ivi/	/iði/	/iwi/	/idi/	/izi/	Total
Original Count	/ɑbɑ/	**63. 0**	1. 0	0	8. 0	0	0	0	0	0	0	0	0	72. 0
	/ɑvɑ/	14. 0	**41. 0**	0	6. 0	4. 0	2. 0	1. 0	4. 0	0	0	0	0	72. 0
	/ɑðɑ/	0	5. 0	**17. 0**	0	3. 0	7. 0	2. 0	1. 0	22. 0	0	15. 0	0	72. 0
	/ɑwɑ/	7. 0	0	0	**48. 0**	0	0	7. 0	0	0	10. 0	0	0	72. 0
	/ɑdɑ/	4. 0	15. 0	13. 0	0	**10. 0**	10. 0	3. 0	7. 0	1. 0	0	7. 0	2. 0	72. 0
	/ɑzɑ/	1. 0	6. 0	13. 0	0	18. 0	**16. 0**	6. 0	0	2. 0	0	10. 0	0	72. 0
	/ibi/	41. 0	1. 0	2. 0	12. 0	0	2. 0	**7. 0**	4. 0	0	2. 0	1. 0	0	72. 0
	/ivi/	6. 0	25. 0	1. 0	0	6. 0	6. 0	0	**19. 0**	1. 0	0	7. 0	1. 0	72. 0
	/iði/	1. 0	2. 0	6. 0	0	3. 0	0	0	2. 0	**29. 0**	0	27. 0	2. 0	72. 0
	/iwi/	0	0	0	0	0	0	3. 0	1. 0	0	**68. 0**	0	0	72. 0
	/idi/	1. 0	2. 0	5. 0	2. 0	6. 0	12. 0	2. 0	4. 0	1. 0	0	**32. 0**	5. 0	72. 0
	/izi/	3. 0	4. 0	1. 0	0	3. 0	2. 0	0	4. 0	1. 0	0	31. 0	**23. 0**	72. 0
	%	87.5	56.94	23.61	66.67	13.89	22.22	9.72	26.39	40.28	94.44	44.44	31.94	

The classification results after merging the VCV sequences for the consonants /d/ and /z/ together are shown in Table 4-20. The percentage of correct classification with the test set in this case was 69.93%. The confusion patterns in the training set between the three classes /ɑðɑ/,/iði/, and classes associated with consonants /z/,/d/ continued in the test set. In addition, classes /ɑbɑ,ibi/ and /ɑvɑ,ivi/ had mutual confusion in both training and testing. The result of testing the

3-class models are presented in Table 4-21. The overall percentage of correct classification was 83.37%.

Table 4-19 Confusion matrix for the 6 class testing phase

Classes		Predicted Group Membership						
		/ɑbɑ,ibi/	/ɑvɑ,ivi/	/ɑðɑ,iði/	/ɑwɑ,iwi/	/ɑdɑ,idi/	/ɑzɑ,izi/	Total
Original Class	/ɑbɑ,ibi/	**117**	8	2	15	0	2	144
	/ɑvɑ,ivi/	21	**87**	2	7	14	13	144
	/ɑðɑ,iði/	3	12	**98**	0	3	28	144
	/ɑwɑ,iwi/	19	1	2	**122**	0	0	144
	/ɑdɑ,idi/	12	15	32	0	**46**	39	144
	/ɑzɑ,izi/	11	28	24	2	31	**48**	144
	%	**81.25**	**60.42**	**68.1**	**84.72**	**31.94**	**33**	

In summary, the pattern of confusions between consonants was similar in both training and testing sets for the different class configurations used in this study. This indicates that the models developed using known utterances can effectively identify unknown utterances. It also indicates that the set of visual features used in this study are capable of representing the VCV sequences used in the study.

Table 4-20 Confusion matrix for the 5 class testing phase

Classes		Predicted Group Membership					
		/ɑbɑ,ibi/	/ɑvɑ,ivi/	/ɑðɑ,iði/	/ɑwɑ,iwi/	/VzV,VdV/	Total
Original Class	/ɑbɑ,ibi/	**114**	7	0	18	5	144
	/ɑvɑ,ivi/	21	**76**	2	8	37	144
	/ɑðɑ,iði/	2	8	**104**	1	29	144
	/ɑwɑ,iwi/	18	0	1	**124**	1	144
	/VzV,VdV/	27	43	56	0	**162**	288
	%	**79.17**	**52.8**	**72.2**	**86.1**	**56.25**	

Table 4-21 Confusion matrix for the 3 class testing phase

Classes		Predicted Group Membership			
		/ɑb,vɑ,ib,vi/	/ɑð, d, zɑ,ið,d,z,i/	/ɑwɑ,iwi/	Total
Original Class	/ɑb,vɑ,ib,vi/	**215**	23	50	288
	/ɑð, d, zɑ/ /ið,d,z,i/	19	**123**	2	144
	/ɑwɑ/,/iwi/	39	4	**389**	432
	%	74.65	85.42	90.04	

Class /ɑb,vɑ,ib,vi/ represent sounds that require complete mouth contact to produce them. Class /ɑwɑ,iwi/ involves sounds that require partial mouth closure in producing them. Both classes were mutually confused in both training and testing set results.

4.1.3 Word-Based Training

The training and testing sets in word-based training are formed by randomly dividing all the word utterances into two equal sized sets. The first half of utterances is used for training and developing the coefficients of the discrimination functions and the second half of utterances is used for testing the coefficients developed in the training process. The LDA functions were calculated for 12-class, 6-class, 5-class, and 3-class configurations. The classification function coefficients with their classification results are shown in Appendix A for the four configurations.

The structural matrix for the word-based training with 12 classes is shown in Table 4-22. SPSS identified the features that have high correlation with the score of each function and those features are marked in bold in Table 4-22. Comparing those features with the features identified by SPSS in speaker-based training with 12 classes shown in bold fonts of Table 4-6 indicates that both training sets had the same features that are highly correlated with the score of the first and second LDA functions. This resemblance of features continued to exist between word-based and speaker based training with 6-classes and 5-classes as indicated in Table 4-10 and Table 4-23 for the 6-class training, Table 4-13 and Table 4-24 for the 5-class training.

Table 4-22 Structural matrix for word-based training with 12 classes

Features	Functions										
	1	2	3	4	5	6	7	8	9	10	11
Slope_LR1	**.580***	**-.534**	.055	.204	.184	.005	-.144	-.185	.267	.364	-.022
LR2	**.548***	-.283	**.426**	.326	.026	-.087	-.249	.216	.190	-.054	-.085
Slope_UP2	**.492***	.174	.054	-.393	-.222	-.003	.342	.303	.207	-.085	.150
Slope_UP1	**-.432***	-.340	-.140	**.408**	.302	-.093	.063	.105	-.245	.079	-.132
Slope_LR2	**-.547**	**.583***	.281	.069	.041	.043	.284	-.042	.014	.157	-.057
LR1	-.215	**.443***	.072	-.109	-.088	.023	-.355	**.412**	.318	-.057	-.196
UP2	**-.406**	-.201	-.219	**.506***	**.405**	.210	.053	.058	.161	-.234	.387
Slope_UL2	-.266	-.242	.197	-.129	-.105	**-.692***	-.227	-.074	.163	-.007	.162
Slope_UL1	.160	.363	-.211	.029	.024	**.442***	-.199	.000	-.134	-.384	-.127
UL1	.097	-.054	.225	**-.414**	**.457**	-.160	**.490***	-.220	-.067	-.021	-.376
LR3	-.046	.353	**.408**	.202	.398	-.181	-.213	**.485***	.048	.070	-.174
UP3	-.021	.054	-.169	.205	.333	.049	.166	.131	.243	-.321	**.659***
UP1	-.033	.101	-.087	.142	.195	.293	-.030	-.026	**.477**	-.252	**.600***
UL3	.023	-.064	.093	**-.420**	.236	-.189	.253	-.336	.254	-.355	**-.534***
UL2	.246	.327	-.209	-.190	.398	.060	.295	-.317	.020	-.288	**-.507***

Table 4-25 shows the structural matrix for word-based training with 3 classes. Comparing the structural matrix for 3 class configurations in word-based training with the structural matrix for 3 class configurations in speaker-based training shown in Table 4-16, the variables that are highly correlated with the classification score of the LDA functions were different from those resulting from the 12, 6, and 5 class configurations.

Table 4-23 Structural matrix for word-based training with 6 classes

Feature	Function				
	1	2	3	4	5
Slope_UP2	**.505***	-.216	-.382	.096	.006
Slope_UP1	**-.464***	.345	**.448**	.219	-.068
Slope_LR1	**.482**	**.640***	.188	-.036	.380
LR2	**.536**	**.616***	-.038	.030	-.094
Slope_LR2	**-.402**	**-.469***	-.129	.197	-.064
LR1	-.159	**-.400***	-.204	-.115	-.100
Slope_UL1	.182	-.393*	.196	-.359	-.005
UP2	**-.433**	.215	**.639***	.089	-.049
UP3	-.022	-.058	.344*	.238	-.147
Slope_UL2	-.270	.263	-.312*	.281	-.237
UP1	-.028	-.085	.239*	-.086	.049
LR3	-.010	-.109	-.098	.340*	-.158
UL3	.013	-.007	-.263	.345	**.599***
UL1	.081	.025	-.286	**.556**	**.563***
UL2	.259	**-.422**	.139	.279	**.545***

Table 4-26 summarizes the classification performance associated with word-based training and speaker-based training for different number of classes. The results of testing the LDA functions developed using word-based training with unknown word utterances were similar to speaker-based testing results. The correct classification percentages for the test set are shown in Table 4-27. The confusion matrices associated with testing word-based functions is included in Appendix A.

Table 4-24 Structural matrix for word-based training with 5 classes

Feature	Function			
	1	2	3	4
Slope_UP2	**-.505***	-.227	**-.403**	.076
Slope_UP1	**.461***	.356	**.441**	.221
Slope_LR1	**-.504**	**.613***	.169	-.037
LR2	**-.548**	**.595***	-.053	-.037
Slope_LR2	**.435**	**-.441***	-.084	.101
UL2	-.245	**-.410***	.141	.288
Slope_UL1	-.168	**-.400***	.248	-.348
LR1	.182	-.399*	-.166	-.185
UP2	**.431**	.226	**.639***	.120
Slope_UL2	.250	.274	-.398*	.351
UP3	.033	-.045	.324*	.273
UP1	.034	-.081	.251*	-.065
UL1	-.078	.036	-.302	**.456***
UL3	-.028	-.004	-.307	.353*
LR3	.044	-.090	-.059	.183*

In a typical classification problem, models are developed based on known parameters; the developed models are tested with parameters coming from unknown sources. The analysis of the results for structural matrices, the classification results of Table 4-26 and the testing results of Table 4-27 suggest that the performance of the discrimination in the training and testing parts is not much affected by the way the data is divided. The speaker based training develops a discrimination model for utterances coming from known speakers, and then this model can be used to classify utterances coming from unknown speakers, which is a configuration that is

closer to a typical classification problem. In the remaining part of this research, tests are performed on models developed by speaker based training.

Table 4-25 Structural matrix for word-based training with 3 classes

Feature	Function	
	1	2
LR2	**.598***	**-.536**
Slope_LR2	**-.458***	**.454**
Slope_UP2	**.454***	.264
Slope_UP1	-.377*	-.306
UP2	-.346*	-.200
Slope_UL2	-.263*	-.244
UL1	.083*	.067
Slope_LR1	**.563**	**-.588***
UL2	.255	**.502***
LR1	-.211	.350*
Slope_UL1	.155	.342*
LR3	-.013	.172*
UP3	.002	.117*
UP1	-.024	.074*
UL3	.006	.052*

Table 4-26 Comparing performance results between speaker based and word based training

Number of Classes	% correct Classification	
	Speaker based Training	word based Training
Twelve	55.3	53.9
Six	65.2	62.3
Five	72.1	71.8
Three	84.4	85.1

Table 4-27 Testing speaker-based and word-based LDA functions

Number of Classes	% correct Classification	
	Speaker-Based testing	Word-Based testing
Twelve	43.1	49.42
Six	59.91	62.15
Five	69.93	71.32
Three	83.37	83.45

4.2 STEP-WISE ANALYSIS

In step-wise analysis, the features are applied to the discrimination problem one at a time. In every step of the analysis, the feature with the highest statistical value in discrimination is entered. This process is continued till all features are applied or the user set thresholds for entering and removing variables are reached. Table 4-28 shows which features were entered in every step of the analysis for different class configurations. The order in which these features enter into the analysis indicates how important those features are in the discrimination problem. Features being admitted to the analysis at later steps have a smaller contribution towards the discrimination problem.

Figure 4-1 shows the effect of adding one feature at a time on the classification performance for different class configuration with the training set. Adding more features in each classifier improves the performance. The increase in the performance becomes stable after adding the 7^{th} feature into the analysis. In the 3 class case, the first two features LR2 and UL2 were enough to obtain high classification score and adding more features resulted in a small drop in the performance.

Table 4-29 shows the training and testing results obtained when the discrimination problem included the top 7 captured in the step-wise analysis. The LDA functions obtained using these seven features were tested to determine the impact of ignoring 8 features on the discrimination problem

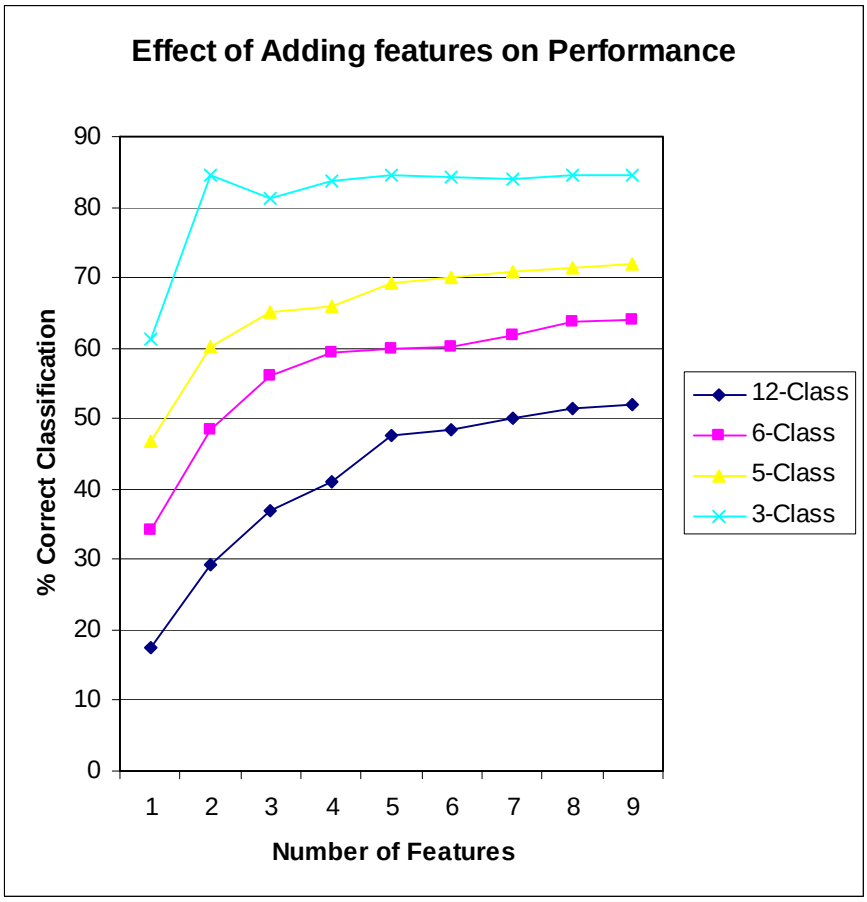

Figure 4-1 Effect of adding features on the classification performance

Comparing these results with the summary of the classification performance when all 15 features are included (Table 4-17), shows that in the training set, these 7 features performed almost as well at the 15 features. The effect of ignoring those features resulted in a 4.8% drop in the training performance for the 12 and 6 class case. The performance difference is almost negligible in the case of training for 5 and 3 classes. In the testing phase, ignoring 8 features resulted in a 3.8% drop for the 12 class case and 1% or less in the remaining classes.

Table 4-28 Features in the order of their importance in different classes

Step number	12 Classes	6 Classes	5 Classes	3 Classes
1	LR2	LR2	LR2	LR2
2	Slope_UP1	Slope_UP1	Slope_UP1	UL2
3	Slope_LR2	UL2	UL2	UL3
4	Slope_UL2	UL3	UL3	Slope_UP2
5	UL2	UP2	Slope_UP2	Slope_UL2
6	UL3	LR1	LR1	LR1
7	UP2	UP3	Slope_UL2	Slope_UP1
8	LR1	Slope_UL2	UP2	Slope_UL1
9	UP1	Slope_LR2	Slope_LR2	UL1
10	Slope_UP2	Slope_UP2	UP3	UP2
11	UP3	Slope_UL1	Slope_UL1	UP1
12	LR3	UL1	UL1	
13	Slope_LR1	UP1	UP1	
14	Slope_UL1	Slope_LR1		
15	UL1			

Table 4-29 Classification performance with the top 7 features

Number of Classes	% correct Classification	
	Training	Testing
Twelve	50.1	39.5
Six	61.9	59.14
Five	70.9	68.68
Three	84.1	83.37

4.3 SPEAKER SPECIFIC DISCRIMINATION

The results presented so far were associated with models that included multiple speakers in their development. Despite the variability across speakers, the models developed were able to classify utterances into different groups. This section shows the results obtained by developing models for each individual speaker, i.e. models that can be used to identify unknown utterances coming from the same speaker.

The data set used in this research consisted of 8 utterances of 12 different VCV words coming from 18 speakers. The utterances for every individual speaker are divided into training and testing sets. The training set consisted of 5 utterances of each VCV word and the testing set consisted of the 3 remaining utterances. Fisher discrimination functions were developed and tested for every speaker. The range of correct classification across speakers in the training and testing phases for different classes are summarized in Table 4-30. The percentage of correct

classification for every speaker for models with different class configurations are presented in figures 4-2 through 4-5.

Table 4-30 Range and average of correct discrimination for 18 speaker-specific models

	# of classes		# of classes		# of classes		# of classes	
	3		5		6		12	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Range of % correct	85-100	55.6-100	88.3-100	50-98.3	80-100	66.7-91.7	90-100	50-86.1
Average % correct	90.06	81.94	93.9	77.49	92.87	75.31	96.29	68.06

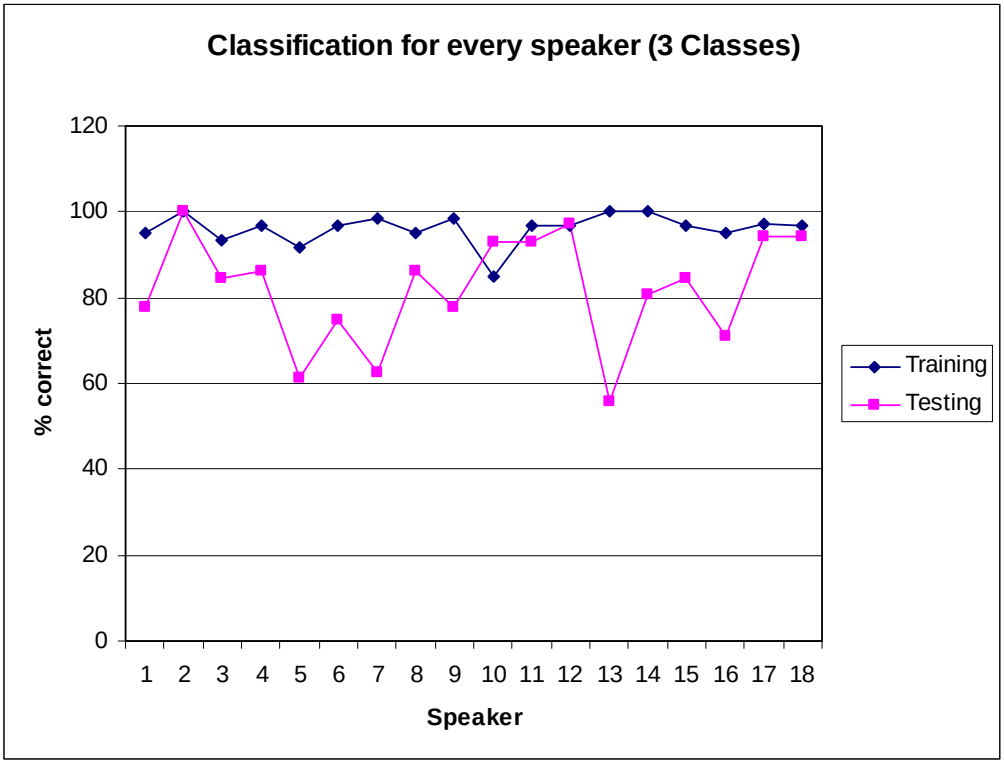

Figure 4-2 Training and testing results for every speaker (3 class configuration)

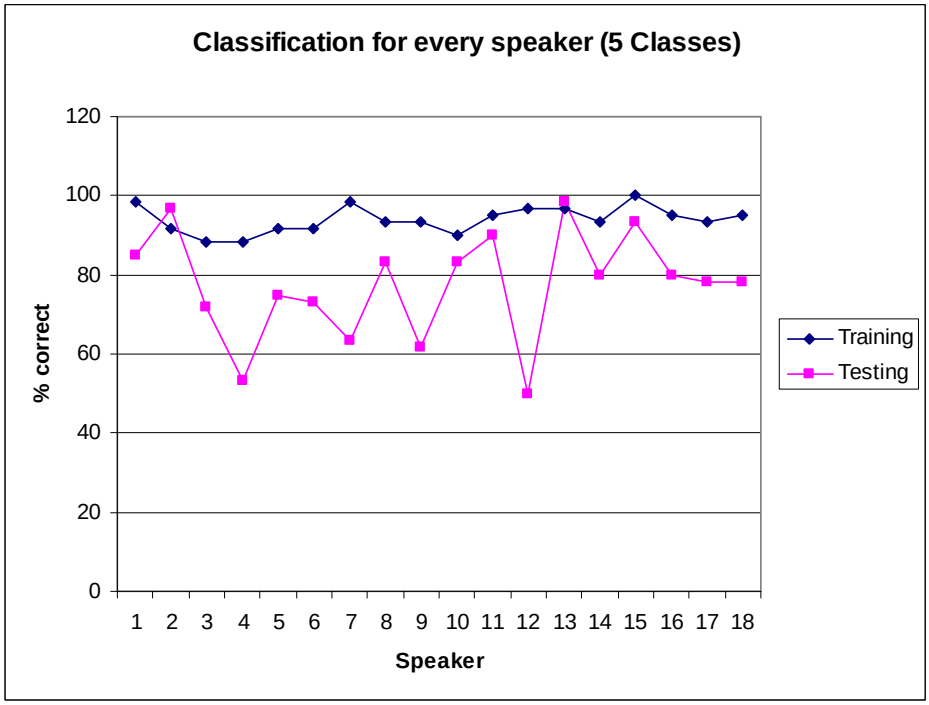

Figure 4-3 Training and testing results for every speaker (5 class configuration)

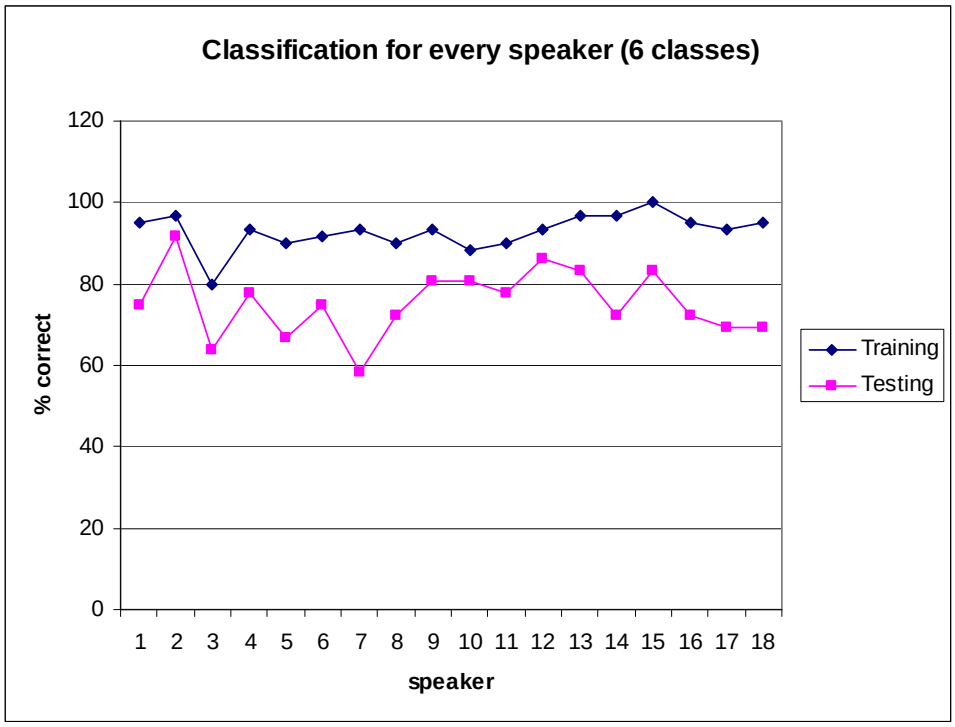

Figure 4-4 Training and testing results for every speaker (6 class configuration)

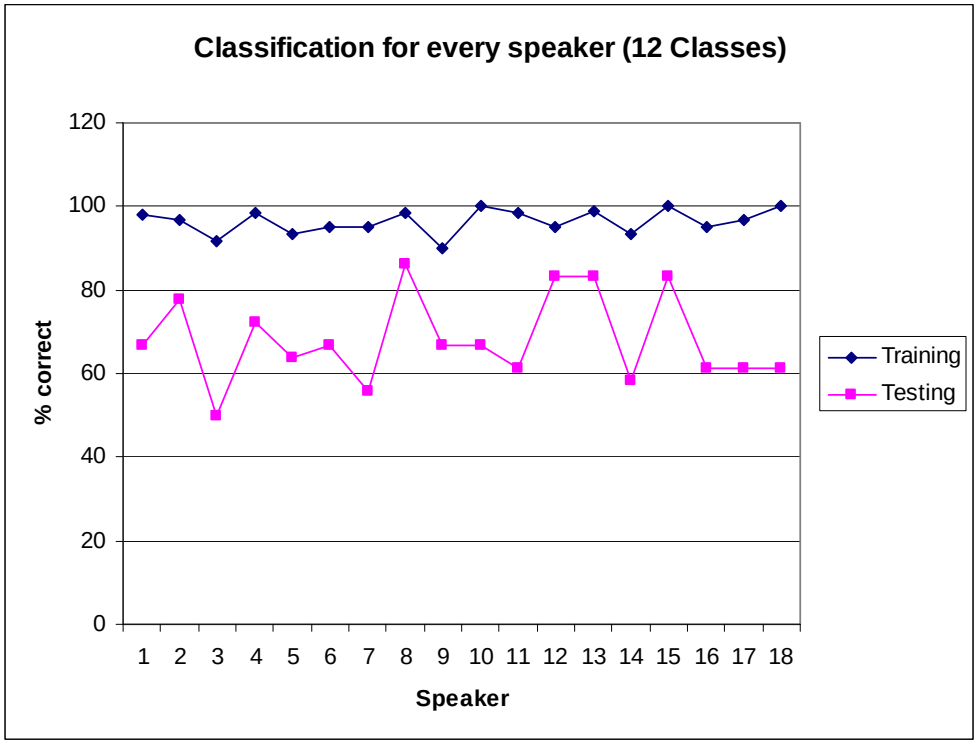

Figure 4-5 Training and testing results for every speaker (12 class configuration)

5.0 DISCUSSION

The objective of this research is to investigate the feasibility of using a small set of features extracted from moving lips to distinguish between the sounds produced. Audio-visual data representing 12 different VCV sounds were obtained from 18 speakers. Several visual features, as shown in Table 3-2, were extracted from the lip motion to represent each of these VCV sounds, and many tests using linear discriminant analysis were conducted on these features to study how well they can classify the different sounds.

5.1 SPEAKER-BASED VERSES WORD-BASED TRAINING

In speaker-based training, the classifier was designed using features from 9 of the 18 speakers. In word-based training, the classifier was designed using features extracted from half of the utterances from all 18 speakers. Each of these configuration schemes models a different problem.

In speaker-based training, a model was developed based on features extracted from known speakers. The resulting model was tested using features from utterances produced by speakers who were not included in the development of the discrimination model. In word-based training, a model was developed based on the contribution of 18 speakers in the data set. The resulting model was tested with unknown utterances coming from the same speakers used to

develop the discrimination model. In this modeling scheme, the training and testing data come from the same source.

The performance of both models on the training data is shown in Table 4-21. The performance of both models on test data is shown in Table 4-22. Results suggest that there is not much difference in the performance of the two. The features highly correlated with the score of the first and second discrimination functions across different classification classes didn't change in both training methods. This indicates that the method of dividing the data didn't result in significant differences in the classification performance.

5.2 PERFORMANCE FOR DIFFERENT NUMBER OF CLASSIFICATION CLASSES

The ideal performance of a classifier would be to distinguish between all 12 VCV words listed in section 3.5. However, Owen's results shown in Table 2-9 suggest that changing the vowel does not affect the classification done visually.

Table 4-4 shows the classification results when the training was done by associating every VCV word to a different class. The average percent of correct classification was 53.3%, and the correct classification percentage for each class was more than that of chance (8%). The highest classification scores were associated with classes /ɑbɑ/ (91%), /ɑwɑ/ (77.8%), and /iwi/ (75%) while the lowest scores were associated with /ɑðɑ/ (27%), /ɑdɑ/ (23%), and /ɑzɑ/ (33%). This indicates that the features used to represent these sounds capture certain characteristics in sounds /ɑbɑ/, /ɑwɑ/ and /iwi/ in a way that distinguishes them well from the other 9 VCV

sounds. However, these features don't perform as well in capturing the difference between the sounds /ɑðɑ/, /ɑdɑ/, and /ɑzɑ/. In addition, the classification scores indicate that the features used in the study were able to detect some difference between the VCV sounds associated with both vowels.

The classification scores and confusions associated with consonants /d/ and /z/ shown in Table 4-4 are consistent with common viseme-to-phoneme mapping of Table 2-11 in which consonants /d/ and /z/ were assigned to the same visual class. Twenty percent of the mis-classified utterances for these sounds are mutually confused, and these two classes have the poorest classification score.

The discussion in section 2.2.2 on the visual perception of phonemes showed that changing the vowel associated with the consonant didn't have an effect on how the observer classified the sequence. To compare more directly to Owen's results, the classification problem was re-defined to assign words with the same consonants to the same class, reducing the number of classes to six classes as shown in Table 4-8. The classification scores for the training data when divided into six classes are shown in Table 4-9. The over all average percent of correct classification is 65.2%.

Merging the vowels resulted in an increase of about 7% in the overall classification performance. As for individual VCV sounds, the sounds with the highest scores continued to be consonants /b/ and /w/ in the 6 class case as they were in the 12 class case. In addition, consonants /d/ and /z/ continued to have the lowest discrimination score in the 6 class configuration as it was in the 12 class. Table 4-9 shows that the discrimination scores for classes associated with /d/ and /z/ were 38.2% and 41.7% respectively. It also shows that 32% of the mis-classified utterances for these sounds were mutually confused. This is consistent with

common viseme-to-phoneme mapping shown in Table 2-9 as discussed in section 4.1.1.2. This led to testing the classification performance when both /d/ and /z/ are considered to belong to the same class, reducing the number of classes to 5 as shown in Table 4-11.

The five-class training results presented in Table 4-12 showed that the class associated with the consonant /b/ and /w/ continued to be detected very well. In addition, merging classes /d/ and /z/ increased the overall percentage of correct classification from 65.2% to 72.1% with individual scores for each class still 3 to 4 times higher than chance (20%). The /d/ and /z/ classes had the poorest performance in the 6 class configuration. When both classes were combined, the recognition for the combined class jumped to 65.3%.

There is still confusion between classes as shown in Table 4-12; Classes /ɑðɑ, iði/, /ɑdɑ, idi/, and /ɑzɑ,izi/. Sixty seven percent of the mis-classified sounds from these sequences were mutually confused. In addition, although class /ɑbɑ,ibi/ and /ɑvɑ,ivi/ involve full closure of lips and have high percentages of correct classification, there is still mutual confusion between both classes. Merging these classes results in a three class problem as shown in Table 4-14.

The three classes presented in Table 4-14 are directly related to the role of the upper and lower lips in producing those sounds. Class 1 represent sounds that involve complete lip contact. Class 2 represents sounds that involve partial lip closure. Class 3 involves sounds with little contribution coming from the lips when producing them. The training results of the 3 class configuration are shown in Table 4-15, with overall performance increasing to 84.4%.

When the number of classes in the training phase was reduced to 3, features highly correlated with the discrimination scores of the first and second LDA functions were (LR2, UL2, UP2, Slope_LR1, Slope_LR2, Slope_UP1, and Slope_UL1). The intuitive meaning for these

features is shown in Table 5-1. Features related to the upper and lower lips became important, while some of the slope features became less correlated with the score of the function.

The features having high correlation with the discrimination score of the first and second discrimination functions were the same for the 12 class, 6 class, and 5 class cases. These seven features were (UP2, LR2, Slope_LR1, Slope_LR2, Slope_UP1, Slope_UL1, and Slope_UP2). These features are related to contributions of the upper lip in the sound production. They capture how fast the upper lip and the lip corners are moving while producing the consonant. The speech production literature discussed in Section 3-3 stated that the upper-lip moves at different rates while producing different sounds. This can explain the significance of slope features at higher number of classes.

The results presented in this section indicate that the features chosen to represent these VCV sounds provide good discrimination between these sounds. The optical reflectors used with the Motion Analyzer could not be used to track the tongue because the adhesive side does not stick to moist surfaces. The tongue plays a role in producing the consonant /ð/ but it has no role in producing the consonants /d/, and /z/. In the context of two dimensional images, the tongue is expected to appear in the image sequence when /ð/ is stated. Capturing this property may reduce the confusion between the classes /ð/ and /d,z/ presented in Table 4-12.

5.3 TESTING THE MODELS

Four different models were developed in the speaker training method. Each of these models had a different number of classes to discriminate between. The Fisher functions for each

model were tested using the utterances coming from 9 speakers that were not used in developing that model, as described earlier in section 3.5.

The results of testing the LDA models for different classes are shown in Table 4-17. The second column shows the classification performance with the training results, and the third column shows the testing results. The testing results are lower than the training results but the difference between the training and testing results decreases as more classes are merged. For the 3 class configuration, the testing and training results are almost identical. For the 12 class case, the drop in the testing results is higher than the drop in other classes. This shows that the automatic classifier performance drops when comparing sequences from two different vowels.

The confusion matrix associated with testing the 12 class case shown in Table 4-24 suggest the following:

- Classes /ɑbɑ/, /ɑwɑ/ and /iwi/ have the highest percentage of correct recognition as was the case with the training data.
- Classes /ɑdɑ/, and /ɑzɑ had low scores, which is consistent with common viseme-to-phoneme mappings that assigned both of them to the same class.
- Class /ibi/ is almost at chance and most of /ibi/ utterances were classified as /ɑbɑ/, i.e. the discrimination functions didn't capture the visual differences between the vowels "ɑ" and "i" for the consonant /b/.

Table 4-20 shows the testing results when the consonants associated with two vowels are assigned to the same class, reducing the number of classes to six. Consonants /d/ and /z/ are mutually confused in testing as they were in training. The results of testing the models after merging these two classes are shown in Table 4-21. The performance of the functions with the testing set was almost equal to the performance with the training set.

The results of testing the model consisting of three classes are shown in Table 4-22. The overall performance of the Fisher functions with the testing data is almost the same as the results obtained using the training data. In addition, the confusion patterns between VCV sequences in the testing set are similar to those of the training set. Furthermore, some of the confusion patterns resulting from the automatic classification are similar to those patterns obtained from visual classification of visemes shown in Table 2-9.

The utterances in the test set came from unknown speakers. Yet, the performance of the developed LDA functions with the unknown data is comparable with the performance of these functions with the training set. This indicates that the developed functions can be used to classify VCV utterances coming from unknown speakers.

5.4 STEPWISE ANALYSIS

The stepwise analysis is a technique to identify which variables contribute more to the discrimination problem. This analysis may help in reducing the dimensionality of the problem by discarding variables with small contribution toward the discrimination. Details of this analysis were discussed in section 3.4.4. and the results of the step-wise analysis for different class configurations are shown in table 4-28.

The intuitive meaning of the visual features used in this study is shown in Table 5-1. The feature LR2 is associated with how much the lip corners close during an utterance. In other words, it reflects the rounding of the lips when the consonant in the VCV sequence is produced. For configurations involving classes greater than 3, the LR2 feature was consistently picked to be the most significant feature in the analysis. The 2nd feature to be picked by the step-wise

analysis was slope_UP1, which represents how fast the upper lip moved towards the lower lip at the beginning of the consonant production in the VCV sequence. When the number of classes involved is higher than 3, slope features representing how fast different points around the lips move in producing a sound become significant. When the training involved 3 classes, the rounding of the mouth (LR2) remained significant; however, the slope information was entered at later stages. The 2nd and 3rd significant features for the 3 class were UL2, and UL3. This is expected since the 3 class case can be classified based on how far the lips close during the production of the consonant (complete, partial or little contribution) as discussed in section 5.2.

The features associated with the production of the consonant are picked up at early stages of the stepwise analysis, while amplitude features related to the 2nd opening to produce the 2nd vowel are picked up at later steps of the analysis. In addition, the slope features become more significant when the number of classification classes increases. This is expected, since the rate at which lips move is different between consonants and for a larger number of classes, the amount of mouth opening is not enough to capture the variation between the visual articulators. This study suggests that the features associated with the consonant in the word (second extrema) play a bigger role in distinguishing between the sounds than the other features.

Table 4-29 shows the results of training and testing new models based on the top seven features appearing in Table 4-28. When comparing the results of developing and testing the LDA models shown with the top seven features with the results shown in Table 4-23 with all 15 features, the percentage of correct classification using 7 features is very close to the corresponding percentage when all features are used. When the LDA functions were developed bases on 12 classes, the percent of correct classification was 55.3% when all 15 features were used. This percentage dropped by 4.8% to 50.1% when only 7 features were used. The

performance difference for different numbers of classes when all features were included in the analysis was better than the performance when 7 features were used by only 2-5% in both training and testing. This suggests that adding features beyond the top seven shown in Table 4-23 results in a small improvement of 2-5% in performance. Reducing the number of features reduces the complexity of the model with minimal loss in classification accuracy.

Table 5-1 Intuitive meaning of the features used in the analysis

Feature	Intuitive Meaning of the Feature in producing the VCV sequence
UL1	How far the upper and lower lips go apart to produce the initial vowel
UL2	How close the upper and lower lips get to produce the consonant
UL3	How far the upper and lower lips go apart to produce the second vowel
LR1	How wide is the mouth when producing the initial vowel
LR2	How wide is the mouth when producing the consonant
LR3	How wide is the mouth when producing the second vowel
UP1	How much the upper lips moved to produce the initial vowel
UP2	How much the upper lips moved to produce the consonant
UP3	How much the upper lips moved to produce the second vowel
Slope_UL1	How fast the upper and lower lips moved in producing the consonant
Slope_UL2	How fast the upper and lower lips moved in producing the second vowel
Slope LR1	How fast the lip corners moved in producing the consonant
Slope_LR2	How fast the lip corners moved in producing the second vowel
Slope_UP1	How fast the upper lips moved in producing the consonant
Slope_UP2	How fast the upper lips moved in producing the second vowel

The top 7 features when the training was done for 12 classes shown in Table 4-23 were LR2, Slope_UP1, Slope_LR2, Slope_UL2, UL2, UL3, and UP2. Some of these features are different from the top seven features that were highly correlated with the discrimination scores of the first and second LDA function of the model (LR2, Slope_UP2, Slope_UP1, UP2, Slope_LR1, Slope_LR2, and LR1). This indicates that a certain feature might contribute towards the discrimination of a specific LDA function, but the same feature might not have significant contribution across all classes. The step-wise analysis evaluates the contribution of features at a more global level than the correlation analysis presented in the structural matrix.

These features provide a new look on existing viseme-to-phoneme mappings. Mappings presented in chapter 2 were based on information observed by the eyes identifying visual representation of sounds, while the features used in this study are simple and easy to extract. The success of the features in achieving good discrimination suggests that they sample the information observed by the eyes of a person experienced in lip reading.

5.5 SPEAKER SPECIFIC DISCRIMINATION

The results associated with the analysis of speaker specific discrimination are shown in Table 4-30, where the average and the range of percentage of correct classification across all 18 speakers are shown.

Some of the speakers scored 100% correct classification in the training and even in the testing set, while some speakers performed poorly, particularly in the testing part. The poor performance in the testing part can be attributed in part to the fact that the training data consisted

of 5 utterances and the testing data consisted of only 3 utterances for every word. Jacknife training and testing might provide better insight on speaker-specific classifiers.

Results presented in Table 4-30 show the overall average performance across speakers in training and testing. The average percentage of correct classification when the training was based on 12 classes was 96.29%. The average for testing with 12 classes was 68.06%. When all speakers were included in the training and testing, the percentage of correct classification was 55.3% in training and 43.1% in testing. The training and testing performances are higher than those results obtained when more than one speaker is involved. This is certainly expected due to smaller within-speaker variability in producing these sounds.

Figures 4-1 through 4-4 show the testing and training results for each of the speakers independently. No specific pattern for the performance of testing the speaker-specific classifiers across different class configurations was observed. Some of the results of testing for speaker-specific models were close to training results, while others were not.

6.0 CONCLUSION

The objective of this study was to investigate the feasibility of utilizing visual cues extracted from lip motion in distinguishing between different sounds. An audio-visual database with 12 VCV sequences was developed. Several features from the lip motion were extracted to represent the VCV sequence. The results suggest that the visual features used provide good discrimination between the VCV sequences used in this study.

Visual features representing the change in mouth shape while producing the consonants in the VCV sequence were the most significant ones. The minimum distance between the upper-lower lips, the distance between lip corners (amount of rounding in the mouth), and rate at which the upper-lip moves while producing the consonant were the features that most effectively captured the differences between the VCV sounds in the study. These features are two dimensional and very easy to extract from video sequences. These visual features manage to capture the uniqueness of the VCV sounds produced by a single speaker at performance rates much higher than the rates when multiple speakers are involved.

The results of this study contribute towards a better understanding of the visemes-to-phoneme mappings and the automatic visual classification of phonemes. Visemes are defined as the visual representation of phonemes. The features used in this study represent a set of parameters that characterize visemes, and they probably represent a sample of what the eye

observes and captures. These features should enable researchers to work on designing automatic classifiers to distinguish between different viseme classes.

7.0 FUTURE WORK

The success in classification of these sounds should encourage attempts to expand the audio-visual data bases to include additional labial English sounds such as /m/, /p/, /f/ ,/θ/, to develop Fisher discriminant functions to identify unknown phonemes. Audio-visual recordings for the new sounds would need to be obtained. Then a lip tracking algorithm similar to the one used by Chen [9, 35] can extract images of the lip-region from successive frames in the video sequence and trace several points around the face in consecutive frames as shown in Figure 3-2. The next step would be to extract the visual features from the distance waveforms for these markers as described in sections 3.2 and 3.3. Grayscale based image segmentation algorithms can be used to detect the appearance of tongue and teeth in each frame in the image sequence. These represent two additional visual features that can be included in the analysis by assigning a binary value of zero or one representing whether lips or teeth were present in a specific frame or not. The extracted features can be used to calculate the Fisher discriminant functions associated with each phoneme. These functions can be used to identify unknown phonemes.

For hearing impaired individuals, lip-reading is not universally well developed and is very limited in new hearing-aid users. This work provides a first step towards building automatic lip reading systems that are based on visual information only. The speaker-specific classifiers discussed in Section 4.4 proved to have good classification. This indicates that the variability of these features for the same word is reduced within one speaker. This becomes important in

situations where a hearing-impaired handicapped person interacts with few people around him or her. In such situations, a visual classifier could be trained using utterances of important words from specific people. Then a signal could be presented to the handicapped person indicating the word to be communicated to him or her. This can contribute towards improving the communication between that hearing-impaired handicap person and the people living with him or her.

Follow-up studies to this work might include studying the relation between the important peaks in the distance waveforms and the acoustic waveform itself. One of the common problems facing people with hearing aids is what is referred to as the cocktail party effect resulting from having more than one person speaking at the same time. In such situations, the noise coming from other speakers occupies the same frequency range as the speaker of interest, which makes it hard to attenuate without affecting the speech signal of interest itself. This study suggests that the features associated with the consonant in the word (second extrema) play a bigger role in distinguishing between the sounds than the other features. Studying the acoustical behaviors of the signal together with the lip waveforms that produced those acoustics may help in identifying instances and points of interest in the acoustical waveform that should be emphasized or de-emphasized. The overall impact of this might improve the intelligibility of speech.

APPENDIX A

WORD-BASED TRAINING

A.1 WORD-BASED TRAINING WITH 12 CLASSES

Table 7-1 Classification Function Coefficients

Features	Functions											
	1	2	3	4	5	6	7	8	9	10	11	12
UL1	1.682	2.118	.351	2.715	.108	.793	1.093	2.373	.388	1.262	3.926	1.923
UL2	-5.974	-7.041	-.076	-6.716	-.063	-2.431	-5.229	-6.164	.353	-4.111	-3.849	-1.042
UL3	5.674	6.095	1.399	6.024	1.534	3.320	5.718	5.300	.717	4.105	3.324	.911
LR1	-4.550	-4.173	-3.657	-1.524	-3.963	-4.148	-3.519	-2.929	-2.670	-1.023	-3.719	-3.896
LR2	1.902	1.911	1.698	-2.843	2.067	2.060	.532	1.223	1.793	1.040	1.943	2.631
LR3	-1.898	-1.496	-1.774	-.468	-1.203	-1.304	.089	-.957	-1.262	-2.251	-.385	-.855
UP1	-.800	-.266	-.752	-.763	-1.295	-.692	-.908	.762	-.180	-.585	.871	-.144
UP2	3.504	-.411	.848	2.217	2.060	.268	.760	-1.226	.657	.540	-1.979	.257
UP3	-2.317	.037	-.289	-1.031	-1.285	.098	.531	.280	-.639	.318	1.222	-.325
Slope_UL1	1.638	24.721	-30.842	6.991	-44.724	-15.928	.965	30.710	-12.520	-2.012	23.586	-3.317
Slope_UL2	-2.981	-26.217	28.585	4.478	72.496	31.513	10.332	-21.481	26.720	31.734	-6.787	37.200
Slope_LR1	-10.649	-15.863	-3.249	-3.723	-13.703	-16.661	1.204	-2.399	-11.180	-34.762	-4.836	-17.544
Slope_LR2	14.093	2.313	3.333	-22.472	1.748	.359	-3.046	3.480	4.336	76.920	10.298	10.672
Slope_UP1	8.239	4.158	-28.823	4.902	-32.880	-18.486	8.300	20.198	-16.681	-1.559	16.472	-4.259
Slope_UP2	39.731	-18.219	6.212	6.109	51.007	6.892	-37.528	-20.083	12.731	-42.798	-23.259	16.368
(Constant)	-22.518	-17.881	-10.803	-20.352	-13.691	-13.820	-17.529	-12.784	-6.520	-19.047	-14.087	-10.024

Table 7-2 Classification results for word-based training with 12-classes

Class		Predicted Group Membership												
		/ɑbɑ/	/ɑvɑ/	/ɑðɑ/	/ɑwɑ/	/ɑdɑ/	/ɑzɑ/	/ibi/	/ivi/	/iði/	/iwi/	/idi/	/izi/	Total
Original Count	/ɑbɑ/	**64**	3	0	2	0	1	2	0	0	0	0	0	72
	/ɑvɑ/	10	**37**	0	2	0	4	2	17	0	0	0	0	72
	/ɑðɑ/	0	5	**33**	0	6	4	1	4	11	0	3	5	72
	/ɑwɑ/	9	0	0	**53**	0	0	2	1	0	7	0	0	72
	/ɑdɑ/	0	5	11	0	**28**	11	3	2	6	0	1	5	72
	/ɑzɑ/	2	7	16	0	16	**12**	1	7	0	0	4	7	72
	/ibi/	19	1	2	5	3	0	**36**	3	0	1	1	1	72
	/ivi/	5	11	0	1	2	5	0	**36**	5	0	3	4	72
	/iði/	0	1	7	0	1	2	0	4	**46**	0	0	11	72
	/iwi/	1	1	1	5	0	0	1	3	1	**59**	0	0	72
	/idi/	1	1	6	0	2	2	0	6	3	0	**35**	16	72
	/izi/	0	0	9	0	11	2	3	3	12	0	5	**27**	72
	%	88.9	51.4	45.8	73.6	38.9	16.7	50.0	50.0	63.9	81.9	48.6	37.5	

A.2 WORD-BASED TRAINING WITH 6 CLASSES

Table 7-3 Classification function coefficients

Feature	Function					
	1	2	3	4	5	6
UL1	1.329	2.062	.242	2.237	1.736	1.218
UL2	-4.586	-5.539	.822	-5.033	-.981	-.923
UL3	4.686	4.702	.428	4.512	1.575	1.354
LR1	-3.341	-2.942	-2.774	-.880	-3.373	-3.550
LR2	1.843	2.134	2.215	-1.002	2.557	2.826
LR3	-.728	-1.086	-1.483	-1.005	-.776	-1.002
UP1	-.397	.577	-.238	-.365	-.010	-.147
UP2	1.170	-1.560	.210	1.062	-.395	-.303
UP3	-.296	.622	-.118	-.257	.219	.223
Slope_UL1	7.603	30.881	-19.132	8.418	-10.004	-6.788
Slope_UL2	11.116	-14.110	33.403	19.726	43.040	41.594
Slope_LR1	-8.826	-13.088	-10.788	-17.774	-13.801	-20.710
Slope_LR2	26.392	21.212	18.028	27.758	22.173	20.425
Slope_UP1	15.747	17.015	-19.590	8.276	-5.990	-7.462
Slope_UP2	-18.022	-34.086	-1.843	-21.331	5.250	.674
(Constant)	-15.592	-11.917	-6.562	-15.675	-10.807	-9.393

Table 7-4 Classification results for word-based training with 6-classes

Classes		Predicted Group Membership						
		/ɑbɑ,ibi/	/ɑvɑ,ivi/	/ɑðɑ,iði/	/ɑwɑ,iwi/	/ɑdɑ,idi/	/ɑzɑ,izi/	Total
Original Class	/ɑbɑ,ibi/	**124**	9	3	5	1	2	144
	/ɑvɑ,ivi/	22	**98**	6	2	6	10	144
	/ɑðɑ,iði/	1	13	**106**	2	9	13	144
	/ɑwɑ,iwi/	24	5	2	**113**	0	0	144
	/ɑdɑ,idi/	5	22	33	1	**55**	28	144
	/ɑzɑ,izi/	6	21	41	0	34	**42**	144
	%	86.1	68.1	73.6	78.5	38.2	29.2	

A.3 WORD-BASED TRAINING WITH 5 CLASSES

Table 7-5 Classification function coefficients

Feature	Function				
	1	2	3	4	5
UL1	1.213	1.857	.113	2.321	1.043
UL2	-4.793	-5.238	.672	-5.171	-1.288
UL3	5.146	4.774	.774	4.733	2.327
LR1	-3.805	-3.254	-3.047	-1.000	-3.822
LR2	1.868	2.210	2.345	-1.242	2.741
LR3	-.639	-1.181	-1.583	-.902	-1.114
UP1	-.484	.427	-.296	-.389	-.289
UP2	1.100	-1.179	.136	1.102	-.311
UP3	-.100	.465	.040	-.316	.423
Slope_UL1	3.756	23.895	-21.662	4.663	-15.047
Slope_UL2	13.344	-2.569	32.435	25.742	44.294
Slope_LR1	-8.844	-14.227	-12.597	-15.901	-19.709
Slope_LR2	26.494	21.700	18.488	26.442	21.022
Slope_UP1	8.628	9.560	-23.201	3.025	-15.201
Slope_UP2	-25.823	-31.348	-8.008	-22.300	-4.432
(Constant)	-16.613	-12.637	-6.967	-16.450	-10.764

Table 7-6 Classification results for word-based training with 5-classes

Classes		Predicted Group Membership					
		/ɑbɑ,ibi/	/ɑvɑ,ivi/	/ɑðɑ,iði/	/ɑwɑ,iwi/	/VzV,VdV/	Total
Original Class	/ɑbɑ,ibi/	**123**	5	0	5	11	144
	/ɑvɑ,ivi/	22	**81**	6	2	33	144
	/ɑðɑ,iði/	1	8	**102**	2	31	144
	/ɑwɑ,iwi/	24	2	2	**113**	3	144
	/VzV,VdV/	11	25	50	1	**201**	288
	%	85.4	56.2	70.8	78.5	69.8	

A.4 WORD-BASED TRAINING WITH 3 CLASSES

Table 7-7 Classification function coefficients

Features	Function		
	1	2	3
UL1	1.269	2.027	.605
UL2	-4.478	-4.742	.247
UL3	4.235	4.146	.754
LR1	-2.770	-.503	-3.028
LR2	1.857	-1.059	2.402
LR3	-1.266	-1.364	-1.309
UP1	.214	-.088	-.202
UP2	-.509	.307	.046
UP3	.203	-.055	-.010
Slope_UL1	19.525	12.996	-15.486
Slope_UL2	-10.866	7.556	36.463
Slope_LR1	-9.374	-17.420	-13.519
Slope_LR2	21.230	24.734	19.052
Slope_UP1	10.829	4.013	-15.530
Slope_UP2	-31.927	-29.103	-.566
(Constant)	-10.791	-13.845	-6.825

Table 7-8 Classification results for word-based training with 3-classes

Classes		Predicted Group Membership			
		/ɑb,vɑ,ib,vi/	/ɑwɑ/,/iwi/	/ɑð, d, zɑ/ /ið,d,z,i/	Total
Original Class	/ɑb,vɑ,ib,vi/	235	8	45	288
	/ɑwɑ/,/iwi/	29	112	3	144
	/ɑð, d, zɑ/ /ið,d,z,i/	41	3	388	432
	%	74.65	85.42	90.04	

APPENDIX B

WORD-BASED TESTING

B.1 WORD BASED TESTING WITH 12-CLASSES

Table 7-9 Classification results for word-based testing with 12 classes

Class		Predicted Group Membership												
		/ɑbɑ/	/ɑvɑ/	/ɑðɑ/	/ɑwɑ/	/ɑdɑ/	/ɑzɑ/	/ibi/	/ivi/	/iði/	/iwi/	/idi/	/izi/	Total
Original Count	/ɑbɑ/	**62**	2	0	1	0	0	7	0	0	0	0	0	72
	/ɑvɑ/	9	**33**	0	4	2	4	3	15	0	0	1	1	72
	/ɑðɑ/	0	5	**28**	0	9	5	4	3	9	0	2	7	72
	/ɑwɑ/	10	0	0	**52**	0	0	3	1	0	6	0	0	72
	/ɑdɑ/	2	3	15	0	**20**	11	4	2	3	0	1	11	72
	/ɑzɑ/	2	8	10	0	15	**8**	4	8	4	0	6	7	72
	/ibi/	24	1	0	5	7	0	**30**	2	0	2	0	1	72
	/ivi/	1	14	1	0	2	3	2	**34**	4	0	7	4	72
	/iði/	1	1	13	0	3	1	0	1	**41**	0	0	11	72
	/iwi/	1	0	1	2	2	0	3	2	0	**61**	0	0	72
	/idi/	1	2	5	0	4	0	0	5	4	0	**32**	19	72
	/izi/	0	0	11	2	2	4	1	6	16	1	3	**26**	72
	%	86.1	45.83	38.89	72. 2	27.78	11.1	41.67	47.2	56.9	84.72	44.4	36.1	49.42

B.2 WORD BASED TESTING WITH 6-CLASSES

Table 7-10 Classification results for word-based testing with 6 classes

Classes		Predicted Group Membership						
		/ɑbɑ,ibi/	/ɑvɑ,ivi/	/ɑðɑ,iði/	/ɑwɑ,iwi/	/ɑdɑ,idi/	/ɑzɑ,izi/	Total
Original Class	/ɑbɑ,ibi/	**127**	5	0	6	0	6	144
	/ɑvɑ,ivi/	21	**95**	4	2	13	9	144
	/ɑðɑ,iði/	4	10	**110**	1	5	14	144
	/ɑwɑ,iwi/	22	5	3	**113**	0	1	144
	/ɑdɑ,idi/	9	21	33	0	**50**	31	144
	/ɑzɑ,izi/	6	22	44	4	26	**42**	144
	%	88.19	65.97	76.39	78.47	34.72	29.17	62.15

B.3 WORD BASED TESTING WITH 5-CLASSES

Table 7-11 Classification results for word-based testing with 5 classes

Classes		Predicted Group Membership					
		/ɑbɑ,ibi/	/ɑvɑ,ivi/	/ɑðɑ,iði/	/ɑwɑ,iwi/	/VzV,VdV/	Total
Original Class	/ɑbɑ,ibi/	**122**	7	0	9	6	144
	/ɑvɑ,ivi/	25	**71**	2	4	42	144
	/ɑðɑ,iði/	2	8	**102**	1	31	144
	/ɑwɑ,iwi/	23	0	1	**119**	1	144
	/VzV,VdV/	18	28	43	0	**199**	288
	%	84.72	49.30	70.83	82.64	69.1	71.32

B.4 WORD BASED TESTING WITH 3-CLASSES

Table 7-12 Classification results for word-based testing with 3 classes

Classes		Predicted Group Membership			
		/ɑb,vɑ,ib,vi/	/ɑwɑ/,/iwi/	/ɑð, d, zɑ/ /ið,d,z,i/	Total
Original Class	/ɑb,vɑ,ib,vi/	**225**	14	49	288
	/ɑwɑ/,/iwi/	25	**118**	1	144
	/ɑð, d, zɑ/ /ið,d,z,i/	39	3	**390**	432
	%	78.12	81.94	90.28	83.45

.

BIBLIOGRAPHY

[1] Sandlin R E *Handbook of Hearing Aid Amplification*, 2nd ed.: Allyn & Bacon 2000.

[2] Summerfield Q, "Lip-reading and Audio-Visual Speech Perception," *Philosophical Transactions: Biological Sciences,* vol. 335, pp. 71-78, 1992.

[3] Sumby W and Pollack I "Visual Contributions to speech intelligibility in noise," *JASA,* vol. 26, pp. 212-215, 1954.

[4] Mcgurk H and Macdonald J, "Hearing Lips and Seeing voices," *Nature,* vol. 264, pp. 746-748, 1976.

[5] Calvert G A, Bullmore E T, Brammer M J, Campbell R, Williams S R, McGuire P K, Woodruff P R, Iversen SD, and David AS, "Activation of Auditory Cortex During Silent Lipreading," *Science,* vol. 276, pp. 593-596, April 25, 1997 1997.

[6] Bernstein L E and Benoit C, "For speech perception by humans or machines, three senses are better than one," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, 1996, pp. 1477-1480 vol.3.

[7] Jialin Z, Chou W, and Petajan E., "Acoustic driven viseme identification for face animation," in *Multimedia Signal Processing, 1997., IEEE First Workshop on*, 1997, pp. 7-12.

[8] Matthews I, Cootes T F, Bangham J A, Cox S, and Harvey R, "Extraction of visual features for lipreading," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 24, pp. 198-213, 2002.

[9] C. Tsuhan and R. R. Rao, "Audio-visual integration in multimodal communication," *Proceedings of the IEEE,* vol. 86, pp. 837-852, 1998.

[10] R. T. Chen, R. R., "Audio-visual integration in multimodal communication," *Proceedings of the IEEE,* vol. 86, pp. 837-852, 1998.

[11] D. Y.-J. C. C. Tsuhan, Dr Simon Lucy, "Audio Visual Speech Data," *Advanced Multimedia Lab at Carnegie Mellon University*.

[12] Fisher C G "Confusions among visually perceived consonants," *Journal of Speech and Hearing Research,* vol. 11, pp. 796-804, 1968.

[13] Chen T, "Audio Visual Speech Processing," *IEEE Signal Processing Magazine,* pp. 9-21, January 2001.

[14] Hans Peter Graf, Erik Casatto, and M. Potamianos, "Robust Recognition of Faces and Facial Features with a Multi-Modal System," *IEEE International Conference on Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation,* vol. III, pp. 2034 -2039, 1997.

[15] Paul Duchnowski, Martin Hunke, and A. Waibel, "Towards Movement Invariant Automatic Lip-reading And Speech recognition," *ICASSP,* vol. I, pp. 109-112, 1995.

[16] Richard Harvey, Iain Mathews, and J. Andrew, "Lip-reading from scale-space measurements," *Proceedings 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 1997.

[17] Bothe H and Rieger F, "Visual Speech and Coarticulations Effects," *ICASSP,* vol. V pp. 634 -637, 1993.

[18] Owens E and Blazek B "Visemes Observed by hearing impaired and normal hearing adult viewers," *Journal of Speech and Hearing Research,* vol. 28, pp. 381-393, September 1985.

[19] Hani Yehia, Philip Rubin, Eric, and Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communications,* vol. 26, pp. 23-43, 1998.

[20] J Barker and F. Bethommier, "Evidence of Correlation between acoustic and visual features of speech," *Proceedings of ICPhS,* pp. 199-202, 1999.

[21] Ezzat T and Poggio T "Miketalk: A talking facial display based on morphing vicemes," *Proc. of Computer Animation,* pp. 96-102, June 1998.

[22] Ezzat T and Poggio T " "Miketalk: A Video Realistic Text-to Audio-Visual Speech Synthesizer," MIT Center for Biological and Computational Learning

[23] Goecke Roland, Millar J Bruce, Zelinsky Alexander, and R.-R. Jordi, "Analysis of audio-video correlation in vowels in Australian English," *In AVSP-2001,* pp. 115-120, 2001.

[24] J.C.Wells, *Longman Pronunciation Dictionary*, Second ed.: Harlow: Pearson Education Limited, 2000.

[25] Ladefoged P, *Vowels and Consonants*, 2nd ed.: Wiley-Blackwell, 2005.

[26] Cohen M and Massaro D W, "Modeling coarticulation in synthetic visual speech," *N. M. Thalmann & D. Thalmann (Eds.) Models and Techniques in Computer Animation, Tokyo Japan,* pp. 139-156, 1993.

[27] Faruquie T A, Kapoor A, Kate R, Rajput N, and Subramaniam L V, "Audio Driven Facial Animation for Audio-Visual Reality," in *Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on*, 2001, pp. 821-824.

[28] Verma A , Rajput N , and Subramaniam L V "Using viseme based acoustic models for speech driven lip synthesis," in *Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 3 (ICME '03) - Volume 03*: IEEE Computer Society, 2003.

[29] Kate Saenko, Trevor Darrell, and J. R. Glass, "Articulatory features for robust visual speech recognition," in *Proceedings of the 6th international conference on Multimodal interfaces* State College, PA, USA: ACM, 2004.

[30] Jintao J, Abeer A, Lynne B , Eduard. T A, and Keating P A, "Similarity structure in perceptual and physical measures for visual consonants across talkers," in *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on*, 2002, pp. I-441-I-444 vol.1.

[31] Jiang Jintao, Alwan Abeer, Auer Edward T , and Bernstein Lynne E "Predicting visual consonant perception from physical measures," *In EUROSPEECH-2001,* pp. 179-182, 2001.

[32] W. M. Salah Werda, Abdelmajid Ben Hamadou, Salah Werda, Walid Mahdi, Abdelmajid Ben Hamadou "Lip Localization and Viseme Classification for Visual Speech Recognition " *International Journal of Computing and Information Sciences,* vol. 4, pp. 62-75, 2006.

[33] Leszczynski M and Skarbek W, "Viseme recognition - a comparative study," in *Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE Conference on*, 2005, pp. 287-292.

[34] W. S. Mariusz Leszczynski, Stanislaw Badura, "Fast Viseme Recognition for Talking Head Application " *ICIAR 2005, LNCS 3656,* pp. 516-523, 2005.

[35] Huang FJ and Chen T "Real-Time Lip-Synch Face Animation Driven by Human Voice," *IEEE Workshop on Multimedia Signal Processing,* December 1998.

[36] Kricos P B and Lesner S, "Differences in visual intelligibility across talkers," *volta review,* vol. 84, pp. 219-225, 1982.

[37] Benguerel M and Pichora-Fuller K, "Coarticulation effects in lipreading " *Journal of Speech and Hearing Research* vol. 25, pp. 600-607, December 1982.

[38] Dodd B and Campbell R, *Hearing by Eye: The Psychology of Lip-reading* London: Lawrence Erlbaum Associates, 1987.

[39] J. Luettin and S. Dupont, "Continuous Audio-Visual Speech Recognition," *Proc. of Fifth European Conference on Computer Vision, Frieburg, Germany,* June 1998.

[40] Petajan E, "Automatic Lip-reading to enhance speech recognition," *Proceedings of IEEE Global Telecommunication Conference,* pp. 265-272, Nov 1984.

[41] B. Yuhus, M. Goldstien, and T. Sejnowski, "Integration of acoustic and visual speech signals using neural network," *IEEE Transactions Speech Audio Processing,* pp. 337-351, 1989.

[42] Nan Li, Shawn Dettmer, and M. Shah, "Lip-reading using Eigenspace," *Proceedgins on workshop on automatic face and gesture recognition,* pp. 30-35, 1995.

[43] Potamianos G and Neti C, "Automatic Speehreading for Impaired Speech," *Proceedings of the Audio Visual Speech Processing Workshop,* Spetember 2001.

[44] R. S. Abdulrauf Biag, and Gilles Vaucher, "A Spatio-Temporal Neural Network Applied to Visual Speech Recognition," *The 2001 IEEE International Symposium on Circuits and Systems,* vol. 2, pp. 329 -332, 2001.

[45] Goldschen A J "Continuous Automatic Speech Recognition by lip-reading." vol. PhD Washington, DC: Goerge Washington University, 1987.

[46] Mase K and Pentland A, "Automatic Lip-reading by Optical flow Analysis," *Systems and Computers in Japan,* vol. 22, pp. 67-75, 1991.

[47] Hiroshi G. Okuno, Yukiko Nakagawa, and H. Kitano, "Incorporating Visual Information into Sound Source Separation," *Proc. of IJCAI-99 Workshop on Computational Auditory Scene Analysis (CASA'99), Stockholm, Sweeden,* 1999.

[48] Cutler R and Davis L "Look Who's Talking: Speaker Detection Using Video and Audio Correlation," *IEEE International Conference on Multimedia and Expo,* vol. 3, pp. 1589-1592, 2000.

[49] Y. H. Takahashi K, "Audio Visual sensor fusion system for intelligent sound sensing," *Proceedings of IEEE International conference on multisensor fusion and integration for intelligent systems (MFI 94)}, Las Vegas, NV,* Oct 2-5 1994.

[50] Ekman P, "Facial Expression and Emotion," *American Psychologist,* vol. 48, pp. 384-392, 1993.

[51] Craig K D, Hyde S A, and Patrick CJ, "Genuine, suppressed and faked facial behaviour during exacerbation of chronic low back pain," *Pain,* vol. 46, pp. 161-171, 1991.

[52] Katsikitis M and Pilowsky I, "A study of facial expressions in Parkinson's disease using a novel microcomputer based method," *Journal of Neurology, Neurosergery, and Psychiatry,* vol. 51, pp. 362-366, 1988.

[53] Ekamn P and Friesen W V, "Facial Action Coding System," *Palo Alto: Consulting Pscychologist Press,* vol. (a), 1978.

[54] Ekamn P and Friesen W V, "Facial Action Coding System:Investigator's Guide," *Palo Alto: Consulting Pscychologist Press,* vol. b, 1978.

[55] Oster H. and Rosentsien, "Baby FACS: Analyzing facial movements in Infants," *Unpublished Manuscript, New York University,* 1993.

[56] Oster Hegely and Nagel, "Adult Adjustment and fine grain analysis of infant facial expressions:Testing the validity of priori coding formulas," *Developmental Pscychology,* vol. 28, pp. 1115-1131, 1992.

[57] Izard C E, "The Maximally Discriminative Facial Movement Coding System," *Unpublished Manuscript, University of Delaware,* 1983.

[58] Essa I A and Pentland A, "A Vision System for Observing and Extracting Facial Action Parameters," *Proceedings of the International Conference on Computer Vision and Pattern Recognition,* pp. 76-83, 1994.

[59] K. a. P. Mase, "Lip Reading by Optical Flow," *IEICE of Japan, Trnasactions,* vol. 6, 796-803.

[60] Yacoob Y and Davis L, "Computing spatio-temporal Representations of Human Faces," *Proceedings In Computer Vision and Pattern Recognition,* pp. 70-70, 1994.

[61] M. Pantic and L. J. M. Rothkrantz, "Expert System for Automatic Analysis of Facial Epxressions," *Image and Vision Computing,* vol. 18, pp. 881-905, March 2000.

[62] ExpertVision, *ExpertVision Operation Manual*. Santa Rosa, CA: Motion Analysis Corporation, 1990.

[63] Godinho Tara, Ingham Roger J, Davidow Jason, and C. John, "The Distribution of Phonated Intervals in the Speech of Individuals Who Stutter," *J Speech Lang Hear Res,* vol. 49, pp. 161-171, February 1, 2006 2006.

[64] Smith A and Kleinow J, "Kinematic Correlates of Speaking Rate Changes in Stuttering and Normally Fluent Adults," *J Speech Lang Hear Res,* vol. 43, pp. 521-536, April 1, 2000 2000.

[65] Wohlert A B and Smith A, "Spatiotemporal Stability of Lip Movements in Older Adult Speakers," *J Speech Lang Hear Res,* vol. 41, pp. 41-50, February 1, 1998 1998.

[66] Stevens K, *Acoustic Phonetics*, First ed.: MIT Press, 2000.

[67] Cohen M, Walker R, and Massaro D "Perception of synthetic visual speech," in *Speechreading by Humans and Machines* New York: Springer, 1996, pp. 153-168.

[68] Gutierrez-Osuna R, "Fisher Discriminant Analysis," in *Pattern Recognition and Intelligent Sensor Machines* College Station: Texas A&M University, Last accessed on June 26, 2008.

[69] Huberty C, *Applied Discriminant Analysis*: John Whiley & Sons, Inc, 1994.

[70] Norusis M J *SPSS 16.0 Statistical Procedures Companion*: Prentice Hall, 2008.