Robust Audio-Visual Person Verification Using Web-Camera Video

by

Daniel Schultz

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

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2006

Author .....
Department of Electrical Engineering and Computer Science
June 29, 2006

Certified by.....
Timothy J. Hazen
Research Scientist, Computer Science and Artificial Intelligence Laboratory
Thesis Supervisor

Certified by.....
James R. Glass
Principal Research Scientist, Computer Science and Artificial Intelligence Laboratory
Thesis Supervisor

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

Robust Audio-Visual Person Verification Using Web-Camera Video

by

Daniel Schultz

Submitted to the Department of Electrical Engineering and Computer Science
on June 29, 2006, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

This thesis examines the challenge of robust audio-visual person verification using data recorded in multiple environments with various lighting conditions, irregular visual backgrounds, and diverse background noise. Audio-visual person verification could prove to be very useful in both physical and logical access control security applications, but only if it can perform well in a variety of environments. This thesis first examines the factors that affect video-only person verification performance, including recording environment, amount of training data, and type of facial feature used. We then combine scores from audio and video verification systems to create a multi-modal verification system and compare its accuracy with that of either single-mode system.

Thesis Supervisor: Timothy J. Hazen
Title: Research Scientist, Computer Science and Artificial Intelligence Laboratory

Thesis Supervisor: James R. Glass
Title: Principal Research Scientist, Computer Science and Artificial Intelligence Laboratory

Acknowledgments

I would like to thank my two outstanding thesis advisors, T.J. Hazen and Jim Glass, for providing me with such a great research opportunity. Their guidance, patience, and understanding has made working on this project a fantastic experience and their encouragement has allowed me to learn a great deal as a researcher. For these and countless other reasons, I am truly grateful to have worked with them.

I would also like to thank everyone in the Spoken Language Systems group. It is a great community that made me feel welcome from the start. Special thanks go out to everyone in the group that helped out with data collection, as I could never have finished this project without your help. I would especially like to thank Kate Saenko for all her assistance with face detection software. She was always gracious enough to answer my questions, no matter how often I asked them.

Finally, I would like to thank my parents, my brother, and my sister for giving me their support throughout the last year, for always listening when I needed someone to talk to, and for making me laugh no matter what else is going on in my life.

This research was supported in part by the Industrial Technology Research Institute and in part by the Intel Corporation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As the number of people entrusting computer systems with their personal and privileged information grows, the secure control of access to that information becomes more and more critical. Current access control methods, such as passwords, are typically lacking in either security or usability, if not both. Speaker verification systems can provide both the high level of security and simple usability which makes them an attractive solution for controlling access to both logical and physical systems.

In order to be effective, the verification system must work in almost any environment with very few errors. When laptop computers were rare, it could be assumed that verification would be run in a relatively quiet, indoor setting with good lighting available. This type of environment is ideal for collecting the audio and video data necessary for verification. As laptops become more common, the list of environmental conditions that speaker verification systems will encounter becomes endless. In order for any computer security system to be effective, it must be tolerant of this environmental variety.

It is important to note that this thesis will focus on techniques for improving person verification systems. Person verification systems are often confused with person identification systems. While similar in many ways, there are subtle differences that are important to understand. Person verification systems begin with some number of known, or enrolled, users. When a user wants to be authenticated, the system is given a sample of audio or video data as well as the credentials of an enrolled user.

The verification system will then determine whether the identity of the speaker in the sample data matches the identity in the given credentials. This is different from a person identification system, in which the system is given only a sample of data and must determine which of the enrolled users is identified in the sample data.

1.1 Motivation

Speaker verification systems can offer advantages over the most common solutions for multiple types of access control systems. For instance, the most common method for controlling access to information in computer systems is the password, a simple string of alphanumeric digits that must be kept secret in order for it to be effective. Despite their popularity, passwords have many flaws that can influence their effective security level. First, the actual security of passwords is generally considered to grow with their complexity. Because of this, strings that include numbers, symbols, and both lowercase and uppercase letters make the best passwords. As users need access to more and more systems, remembering such a password for each of these systems becomes difficult. This can cause a user to apply the same password to multiple systems or to use simpler passwords that provide little actual security. Also, passwords can be easily stolen. If an imposter can find an enrolled user's password, any security offered by the password will be erased.

The usefulness of speaker verification systems is not limited to logical access control. It can also offer advantages over the most common method for physical access control, the key. Unfortunately, keys, like passwords, have many flaws. First, in most cases, a key only works for one lock or one location, and therefore a user must carry a key for each lock that they would like to be able to open. Second, keys, like passwords, can be easily stolen and when this occurs, any security provided by the lock and key disappears.

A speaker verification system can solve the problems of both these systems. With a speaker verification system, there is no secret to remember. A user's face or voice becomes their password or key and it can be used for any number of systems without

decreasing the level of security. This is because, unlike a password, it would be very difficult to copy someone's face or voice.

The one advantage that keys and passwords have over speaker verification is that they will always work. So long as the user remembers their password for a particular system or has the correct key for a lock, he or she will always be granted access. With speaker verification systems, errors can occur which can let in imposters or lock out enrolled users. Reducing this error rate is essential for making speaker verification systems a viable solution for security systems.

1.2 Previous Work

Much research has been done on multi-modal speaker verification and closely related topics. However, there is little research that has taken on exactly the same challenge that this thesis deals with. For instance, the Extended Multi Modal Verification for Teleservices and Security (XM2VTS) project [3] has built a large publicly-available database containing audio and video recordings to be used for multi-modal person identification or verification system experiments. However, the subjects in the database were all recorded in a highly controlled environment with a solid blue backdrop. The environment for recording audio, which was recorded with a clip on microphone attached to the subject, was also highly controlled to keep the noise level at an absolute minimum. Using such a controlled environment is an unrealistic test for a speaker verification system which is likely to experience widely varying lighting and noise conditions. This is especially true for a verification system that is to be used on mobile devices.

The work by Ben-Yacoub et al. described in [4] performed person verification, but the fact that it uses the XM2VTS database's controlled environment recordings puts it in a separate category from the work in this thesis. Research explained by Fox and Reilly in [5] also uses the controlled environment recordings from the XM2VTS database, but that is not the only thing that sets it apart from the work described in this thesis. The work done by Fox and Reilly was in speaker identification, whereas

this thesis investigates multi-modal speaker verification. The important distinction between identification and verification was described earlier in this chapter.

The work done by Maison et al. in [10] is also focused on speaker identification. However, unlike the work of Fox and Reilly, this work uses data from broadcast news, providing a large number of background noise and lighting conditions. While the environmental variation is similar to the work in this thesis, our speaker verification system must work with much more limited enrollment data sets. In a system designed to identify broadcast news anchors, there is a wealth of audio-visual data for each enrolled news anchor. In order for our verification system to be practical, new users must be added with short enrollment sessions. Such limited data sets for enrolled users can decrease the robustness of the verification system.

Finally, the speaker identification work done by Hazen et al. in [8] and [7] shares many characteristics with the work performed for this thesis. Recordings were performed on comparable hardware, using a camera on a handheld to record images. Their work used audio recordings of single phrases along with face images. However, neither of these works used video recordings. They instead used single-image snapshots of the face, whereas this thesis uses video. While the environment was much less controlled in these works than in the research performed on the XM2VTS data, the environments in this thesis are even more varied than in these two papers. In addition to the office recordings, we recorded in a busy cafe and near a heavily-trafficked street. Also, our video frames have a resolution of 160x120, whereas this earlier work uses 640x480 resolution frames.

1.3 Goals

This thesis attempts to build a robust multi-modal person verification system. Verification, not identification, is the task that such a system would perform if it is to be used for access control. We also want to build a system that is capable of operating effectively in a wide variety of uncontrolled environments. It is not always possible for authentication to be performed in a well-lit, noiseless location with a carefully chosen

background. The verification system must be able to handle changes in environment gracefully. Finally, we also want the verification system to operate well with limited enrollment data. It is inconvenient for users to be required to record a large amount of data in order for verification to work properly. We hope to provide robust person verification that will perform effectively in spite of these challenges.

1.4 Outline

The rest of this document is organized as follows:

- Chapter 2 describes the process by which data was collected, provides information on the subjects recorded, and discusses the quality of the audio-visual data. Chapter 2 also describes the utterances that were recorded and the locations where data collection was performed.
- Chapter 3 explains the progression from raw recorded video to processed video used for experiments. This includes the video format conversion, face detection, and face recognition steps.
- Chapter 4 explores the effects that several factors, when taken individually, will have on speaker verification performance. These factors include the location where video is recorded, the inclusion of in-set imposters in testing, the size of the data sets used for training, the criteria by which training data is selected, and the particular facial features used.
- Chapter 5 explains the process for training and testing the audio verification system.
- Chapter 6 describes the results of combining an audio speaker verification system with a visual person verification system to create a multi-modal person verification system. This chapter compares the effect that location has on a multi-modal system with its effect on either single-mode verification system.
- Chapter 7 summarizes results and proposes future work.

Chapter 2

Data Collection

In order to test the effectiveness of different techniques for audio-visual person verification, we first needed to collect a set of audio-visual data. We recorded subjects while they read a list of utterances, which were either short sentences or strings of digits. Data was collected using a Logitech QuickCam Pro web-camera attached to a laptop. The portability of the laptop allowed us to collect data in multiple environments. As a result of recording in multiple environments, the data set contains a variety of noise levels and lighting conditions.

2.1 Recording Locations

During each session, a subject was recorded in three separate locations. The first location was a quiet, well-lit office setting. In this location, there was generally very little noise and the lighting conditions were very consistent from one subject to the next and from one day to the next.

The second location was on the first floor of an academic building which contains lecture halls and a cafe. During recording times, this location often had high amounts of foot traffic. Because of the high traffic of this location, there was often a great deal of crowd noise. Also, when recording near the cafe, there was a variety of sounds that one would normally associate with a busy restaurant. This noise can vary greatly from one video to the next or from one recording session to the next. Additionally,

the lighting conditions in this location were much less consistent than the lighting of the office setting. There was a mix of natural and artificial lighting that varied greatly across different areas of the first floor having a noticeable affect on video quality.

Finally, each subject was recorded in an outdoor setting near a busy intersection. The intersection contained heavy motor traffic during the day, including a great deal of traffic from large trucks. The rumble of engines as well as the sounds of sirens from police cars and fire engines can be heard in some videos recorded in the outdoor location. Another ingredient of the noise in the outdoor videos was the wind. There was noise from wind blowing directly into the microphone as well as wind rustling the leaves of nearby trees. Lighting in the outdoor recordings also varied the most of the three locations. There was a drastic difference in lighting based on whether it was a sunny day, a cloudy day, or a rainy day. Also, if the subject was facing the sun the lighting quality could be excellent, whereas with his or her back to the sun, the subject's face was often very dark with little contrast.

An example image from each of the three locations can be seen in Figures 2-1, 2-2, and 2-3.

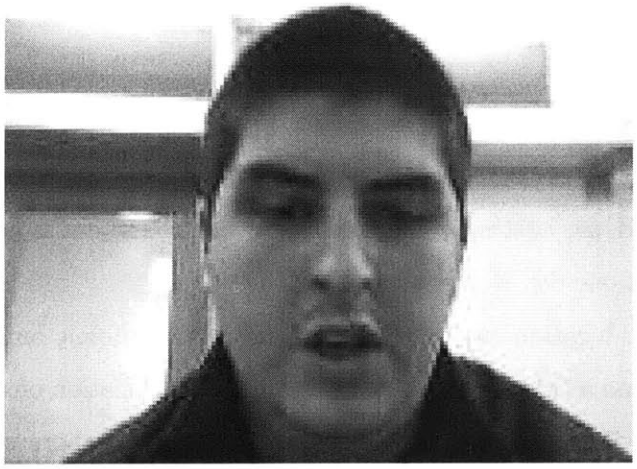

Figure 2-1: Example Frame from the Office Environment

Figure 2-2: Example Frame from the Cafe Environment

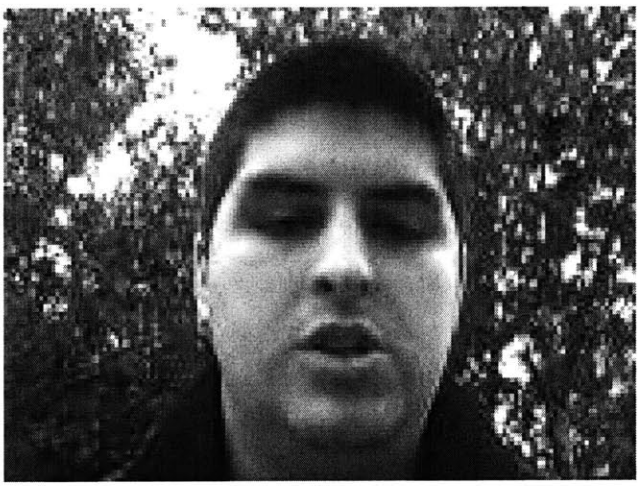

Figure 2-3: Example Frame from the Street Environment

2.2 Recording Protocol

For each recording session, the subject read the same eleven utterances in each of the three locations. In each location, the subject was given a place to sit and did not move from this position until he or she completed the recordings for the location. The recording sessions always began in the office setting. The subject would start recording, read one utterance, and then stop recording. This way each utterance was recorded to a separate video. After the subject recorded each of the eleven utterances in the office setting, he or she repeated the recordings in the downstairs cafe setting, followed by the outside setting. Once complete, each session contained 33 total recordings. Subjects were allowed to re-record any utterance if they were unhappy with the previous recording or if they misread the utterance.

2.3 Utterances

The utterances being read were either short sentences or strings of digits. The subjects read from two lists of utterances. The first time a subject did a recording session, he or she read the utterances from the list in Table 2.1, and if he or she returned to do a second recording session, he or she recorded the utterances from the list in Table 2.2. The digits utterances are all composed of digits from one to nine. To provide consistency within the recordings, the digit zero was left out because people read it in different ways. Some people read zeros as "zero," while others read them as they would read the letter "O."

2.4 Video Specifications

The video was recorded in 24-bit color at a resolution of 160 pixels by 120 pixels. While the frame rate of the video varied, the video was typically between 25 and 30 frames per second. The length of the videos also varied but was usually between 4 and 8 seconds.

1	She had your dark suit in greasy wash water all year.
2	Don't ask me to carry an oily rag like that.
3	1 5 7 2 4 6 8 9 2 1
4	2 9 4 4 3 8 7 1 7 6
5	3 6 2 7 3 1 8 4 9 7
6	4 1 3 3 9 5 2 2 6 5
7	5 5 4 7 8 1 4 2 3 4
8	6 9 9 3 5 8 2 5 1 9
9	7 4 5 3 7 5 9 1 1 2
10	8 8 5 6 6 7 7 9 8 3
11	9 6 3 2 8 6 1 6 4 8

Table 2.1: Utterances for Recording Session One

1	She had your dark suit in greasy wash water all year.
2	Don't ask me to carry an oily rag like that.
3	1 5 3 4 3 9 5 9 3 2
4	2 3 8 7 5 4 8 8 4 7
5	3 6 9 6 8 3 3 5 5 6
6	4 2 5 8 1 8 6 6 2 8
7	5 1 7 2 4 5 7 3 1 1
8	6 7 6 5 2 6 4 1 9 4
9	7 4 4 9 9 2 9 7 8 5
10	8 9 1 3 7 7 1 4 6 3
11	9 8 2 1 6 1 2 2 7 9

Table 2.2: Utterances for Recording Session Two

2.5 Subject Statistics

We recorded 100 subjects in total. Fifty of these subjects returned to do a second recording. The subjects with two recording sessions became the enrolled users, while the subjects with only one recording session became the imposters.

The only requirements for the subjects were that they were over eighteen years old and could speak English fluently. Of the 100 subjects, 41 were male, 59 were female. There were 86 native speakers of American English. Of the 14 non-native speakers, many were native speakers of Chinese, though there were also subjects whose native language was UK English, Dutch, and Gujarati, an Indian dialect.

Chapter 3

Video Processing

There are many steps required to process the recorded video into experimental results. The video must first be converted to a format that is readable by face detection software. Face detection must then be run on the video. Face or facial feature images must be extracted from individual frames. Then the video-only verification system must be trained before being tested with experimental data.

3.1 Decompressing Recorded Videos

The videos recorded with the web-camera were stored in a compressed AVI format. In order for the face-detection software to read the video, they needed to be in an uncompressed format, so that video frames could be read individually. Using the QuickTime libraries for Java, we converted the videos from their compressed AVI format to an uncompressed format. We also removed the audio from the clips at this time, since the audio was processed separately from the video. Finally, we down-sampled the video from 24-bit color to 8-bit grayscale. This kept the data set at a manageable size. Using full-color images would have required an amount of time and computing resources that were not reasonable for our experiments. Even after reducing the size of the data sets, single experiments frequently took a computer nearly a day to complete.

3.2 Face Detection and Feature Extraction

After the video was converted to uncompressed, grayscale AVI format, we ran each video through face detection software written in MATLAB [13]. The software we used utilizes the Open Computer Vision software package originally developed by the Intel Corporation [1]. The OpenCV libraries use the face detector developed by Viola and Jones. The algorithms used for their face detector are explained in [14]. The face detection software attempted to determine several key facial features in each frame of video. These features included the center and width of the face and the center, width, and height of the mouth. The values for each of these variables was recorded for use in the feature extraction phase of video processing.

In Figure 3-1, a sample frame is shown on the left. The rectangle superimposed on the face displays what the face detection software determined to be the approximate boundary of the lips. The circle in the middle of the rectangle is the approximate center of the lips. The face detection software also determined the boundary of the face for this frame. The image on the right in the figure shows the loosely-bounded face image that was extracted for this frame.

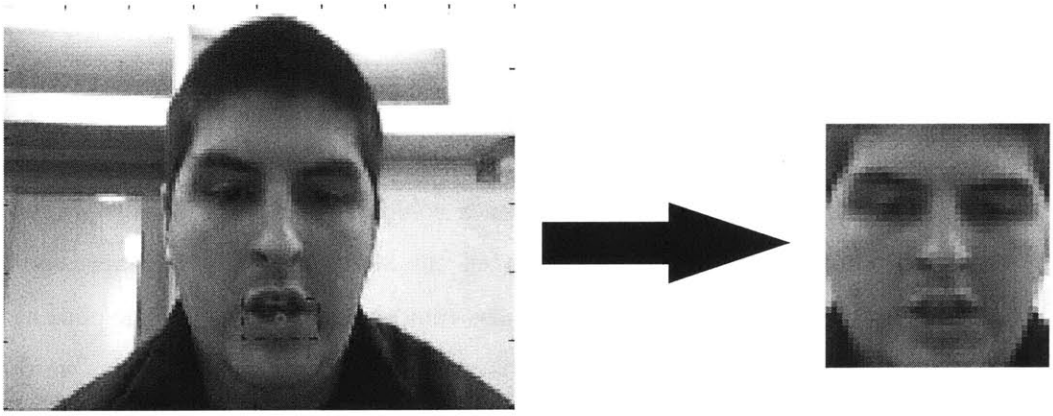

Figure 3-1: Single Frame and Extracted Face Image

Once the face and facial features were detected for each frame of video, images of a face or facial feature were extracted.

3.3 Training the Verification System

Once the videos were reduced to batches of face or facial feature images, a verification system was trained. For our experiments, we trained support vector machines to produce verification scores which we used to test the effectiveness of various techniques for person verification. We used both global and component-based methods for face verification similar to the methods used in [9]. We used the support vector machine package SvmFu [2] for visual speaker verification. The support vector machines were trained in a one versus all manner. This means that each support vector machine was trained to recognize exactly one of the enrolled users. Therefore, there were 50 support vector machines total, one for each enrolled user.

The support vector machines were trained using the images from the second session of each of the 50 enrolled users. If the images are from the one enrolled user the support vector machine is trying to recognize, those images were used as positive training examples. The images from the other 49 enrolled users were the negative training examples. No imposter data was used to train the support vector machines. Training the support vector machines using imposter data would introduce a bias as explained in Section 4.5.

3.4 Testing the Verification System

After training was completed using the images from the second session of the 50 enrolled users, the verification system was tested using the images from the first session of the 50 enrolled users and 50 imposters as the test data.

The testing data was comprised of the images from the 50 imposter recording sessions as well as the images from the first recording session of the 50 enrolled users. Each of the 50 support vector machines produced a verification score for each image from the testing data set. The returned score represented how closely the input image matched the training data for that support vector machine. The higher the score, the closer the image resembled the data from the training set.

It should be noted that in all of the experiments, we evaluated the effectiveness of different techniques based on the equal error rates of each tested technique. To determine the equal error rate, a threshold value is set. Images with scores above this threshold value are accepted, while images with scores below the threshold are rejected. The equal error rate is determined by examining two probabilities: the miss probability and the false alarm probability. The miss, or false rejection, probability is the probability that given some threshold value, the system will incorrectly reject an enrolled user. The false alarm, or false acceptance, probability is the probability that given some threshold value, the system will incorrectly accept an imposter. Given a list of scores from imposters and enrolled users, we adjust the threshold value until it yields the same value for the miss probability and false alarm probability. When each of the two probabilities have the same value, that value is the equal error rate.

Chapter 4

Video Experiments

Before attempting to combine the results of separate audio and video verification systems into a multi-modal person verification system, we needed to examine the effects of different video techniques taken individually. First, we looked at the effect of scoring frames individually versus giving one score for each whole video. Next, we examined how the equal error rate of person verification is effected by the size of the training set. We also tested the effectiveness of randomly selecting training images as opposed to simply selecting images from the beginning of each video. Then we examined the effects of using a training set of images from the recording location that matched their testing data, did not match the testing data, or used a training set that was a mix of all three recording locations. We then looked at the effect of testing the verification system with in-set versus out-of-set imposters. Finally, we experimented with using different facial features and performed some preliminary experiments using combinations of the individual features.

The experiments will often refer to two distinct sets of data. The training data set is the set of all videos from the second recording session of each of the enrolled users. The testing data set included the videos from the recording sessions of the imposters as well as the first recording session from each of the enrolled users. Imposter data was never used to train the verification system.

4.1 Frame Scores Versus Video Scores

For the first video-only experiment, we wanted to compare the equal error rate achieved when frames are treated independently with the equal error rate achieved by combining frame scores in simple ways to create one score per video. We began by testing the verification system using individual frames. We trained the support vector machines using 1000 face images from each of the 50 enrolled subjects in the training data set. Furthermore, we only used images from videos recorded in the office setting. To test the system, we used all of the face images from the testing data set videos recorded in the office setting. Each of the testing images was input to each of the 50 support vector machines and the scores recorded.

We tried two simple techniques for combining individual frame scores to create a per video score. The first technique was an average of each of the frame scores in each video. If frame scores are treated independently, a single frame could produce an outlier score that is high enough to be above the acceptance threshold. If this were the case, an imposter could be verified incorrectly by the system. By the same token, a single outlier could also be enough for an enrolled user to be rejected if the score for that frame was below the threshold. By averaging the frames, the influence of a single frame on the overall score is reduced by the number of frames in the video.

The second technique that we tried was to use only the maximum frame score for each video. For each video, only the best frame score returned by each of the fifty support vector machines was kept. All other scores were discarded.

After computing these three sets of data, the results were used to create a detection error tradeoff (DET) curve for each set. The three curves are shown in Figure 4-1

The detection error tradeoff curves for each of the three sets of results are similar. The frame score approach slightly outperformed each of the per video score approaches, but the equal error rates for the three methods were very similar. Frame scores achieved an equal error rate of 10.18%. The maximum video score approach was next with 10.53%. Finally, the average video score method yielded an equal error rate of 11.21%.

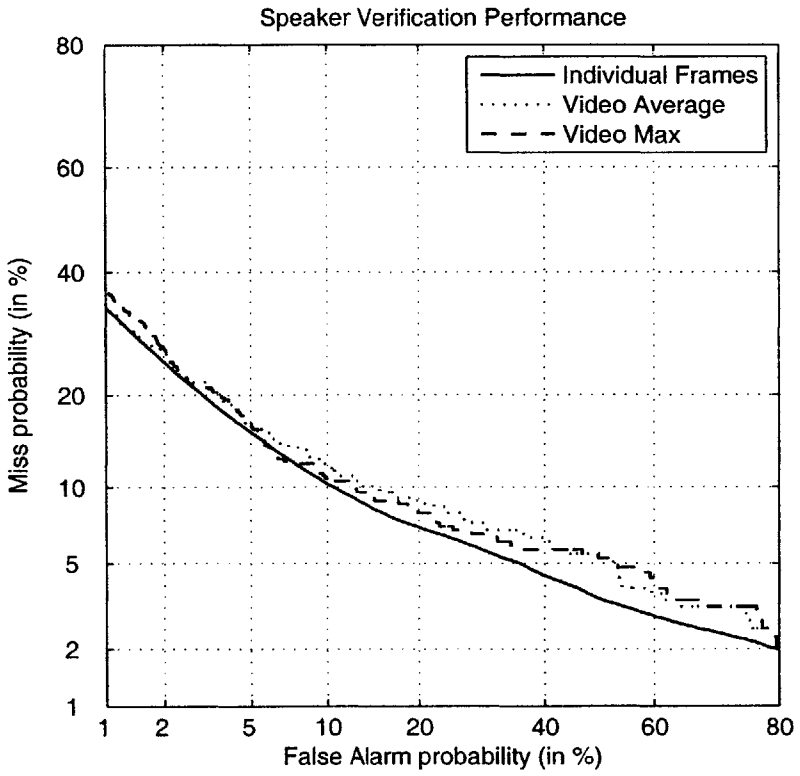

Figure 4-1: DET Curves for Single-Frame, Video Average, and Video Max

4.2 Size of Training Set

Another aspect of the verification system that we wanted to examine was the effect that the size of the training set had on the equal error rate of the system. There are practical reasons for understanding the effect of training set size on the system's equal error rate before continuing with later experiments. The biggest incentive for this experiment was to determine if smaller training sets could be used to produce similar, if not better, equal error rates than a larger training set. Since computational cost grows with the size of the training set, if some reduction in training set size could yield at least similar results, it would reduce the time needed to perform subsequent verification experiments without sacrificing the quality of the results.

For this experiment, we trained the support vector machines using face images from only the office setting from the training set. We performed three trials with training sets that used 100, 500, and 1000 images per subject. In each trial we ran the complete set of face images from the office setting of the testing data and recorded the results. The detection error tradeoff curves for each of the three trials can be seen in Figure 4-2.

The equal error rate was 14.42% when we used 100 images per subject, 11.22% when we used 500 images per subject, and 10.18% when we used 1000 images per subject. While there is a significant drop in equal error rate when the training set size was increased from 500 to 1000 images per subject, the difference in equal error rate was smaller than when increasing the training set size from 100 to 500. Doubling the size of the training set only reduced the equal error rate by about one percentage point. The one percentage point is significant enough that we decided to use 1000 images per subject in all subsequent experiments. However, in order to keep the required processing time at a reasonable level, we decided not to use more than 1000 images per subject in the training set.

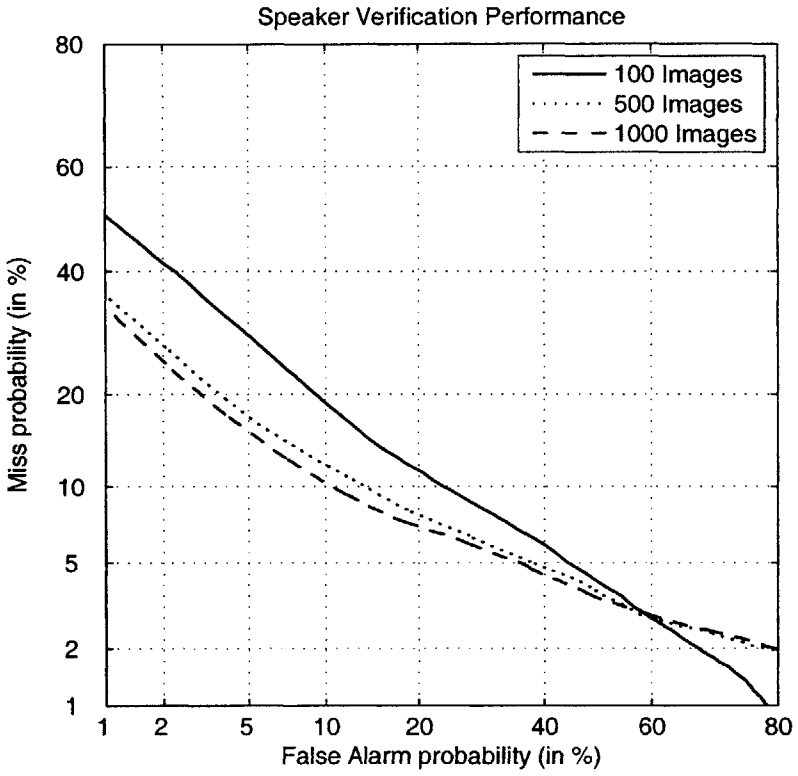

Figure 4-2: DET Curves for 100, 500, and 1000 Training Images

4.3 Matched, Mismatched, and Mixed Training Sets

For this experiment, we wanted to examine the effect that the recording environment has on the effectiveness of the verification system. Since we recorded in three different environments–the office, the downstairs cafe, and the outside street–we could try different combinations of choosing a training environment and testing environment. Because we can also train with data from multiple environments, we essentially have a fourth set of data to train with: the mixed-environment training set.

With four environments to use for training data and three environments to use for testing data, we had twelve different experiments. Each of these experiments fell into one of three categories. The first category was the matched case and included experiments whose training and testing data both came from the same environment. The second category included all experiments using a single-environment training data set that did not match the testing data set. Finally, there was the mixed category which included the three experiments in which mixed-environment data was used for training.

We ran all twelve experiments using face images. For the mixed-environment experiments, we trained the support vector machines using 40 images from each video in the three environments. Since there are 27 videos of digits for each subject, the mixed-environment training set could include up to 1080 images, whereas the single-environment cases all used 1000 images. Based on the results from Section 4.2, the 80 extra images were not likely to provide a significant advantage to the mixed-environment experiments.

		Testing Environment		
Training Environment		Office	Cafe	Street
	Office	10.18 %	23.69 %	16.99 %
	Cafe	24.94 %	22.28 %	23.24 %
	Street	26.42 %	24.18 %	21.57 %
	Mixed	10.21 %	13.14 %	11.82 %

Table 4.1: Equal Error Rates for Video-Only Training/Testing Environment Pairs

The results of the twelve experiments can be seen in Table 4.1. These experiments yielded many interesting results. First, for each set of single-environment training sets, the best equal error rate came from the matched experiment. Given some environment that the verification system was trained with, the best environment for that verification system to test on is the environment that it was trained with. However, when given a testing environment, the best results do not always come from the matched case. Testing in the office or cafe settings is most accurate when using matched training data. However, testing in the street condition is more accurate using office training data than it is with street training data.

Another interesting result from these experiments was that for any environment used for the testing set, training the support vector machines with the mixed-environment data produced equal error rates that were as good or better than any experiment using single-environment training data. The mixed-environment trained support vector machines were able to verify speakers in the office testing set almost exactly as well as when the support vector machines were trained on only office data. For systems tested with data from the cafe or street settings, the mixed-environment system drastically outperformed the single-environment systems at verification. Because the mixed-environment training data contained more variety in lighting, the support vector machines were more likely to find the parts of the images that were consistent across the three environments of the training data. Since the lighting was not consistent across all the training images, it should have been easier for the support vector machines to reduce or eliminate the effects of lighting on the training images and, therefore, look for features of the face and not features of the lighting in a particular environment.

4.4 Consecutive or Random Image Selection

Since each subject that we recorded for these experiments controlled when the recordings started and stopped and since each subject read at a different speed, the number of frames for each video varied greatly. Some videos have less than 100 frames, while

others have several hundred. Since nearly all of the training sets had well over 1000 images to choose from for each subject, we selected 1000 images to use to train the support vector machines. For most experiments, we simply chose the first 1000 images starting from the first recording. Because the environment can change from one video to the next, we thought there could be an advantage to training with images sampled from all the videos. Training with all the videos could help the support vector machine determine which aspects of the frames were characteristic of the individual and which were characteristic of the environment. As was demonstrated with the mixed-environment training experiments in Section 4.3, a greater variety in the training images can lead to lower equal error rates.

In this experiment, we tested a verification system that was trained using the first 1000 face images from each subject in the office setting. We also tested a system that was trained using 1000 randomly selected face images from the office setting. In both cases, the system was tested using all of the images from the office setting in the testing data. The detection error tradeoff curves can be seen in Figure 4-3.

The system achieved an equal error rate of 9.87% when randomly selecting the training images versus 10.18% when simply using the first 1000 images from each subject. While this is only a modest improvement in equal error rate, it is promising that such a small change in the way the system is trained can produce noticeable results. At the very least it serves to verify the results from Section 4.3 that more variety in the training environment can produce more accurate verification systems.

4.5 In-Set Versus Out-Of-Set Imposters

In all of the experiments, the support vector machines were trained using the data from one recording session each for the 50 enrolled users. The testing set contains images from the second recording session of each enrolled user as well as the recording session of each imposter. In this testing scenario, the subjects whose images are used to test the verification system fall into one of three categories. The first is the correct user. This is the person that the support vector machine was trained to recognize.

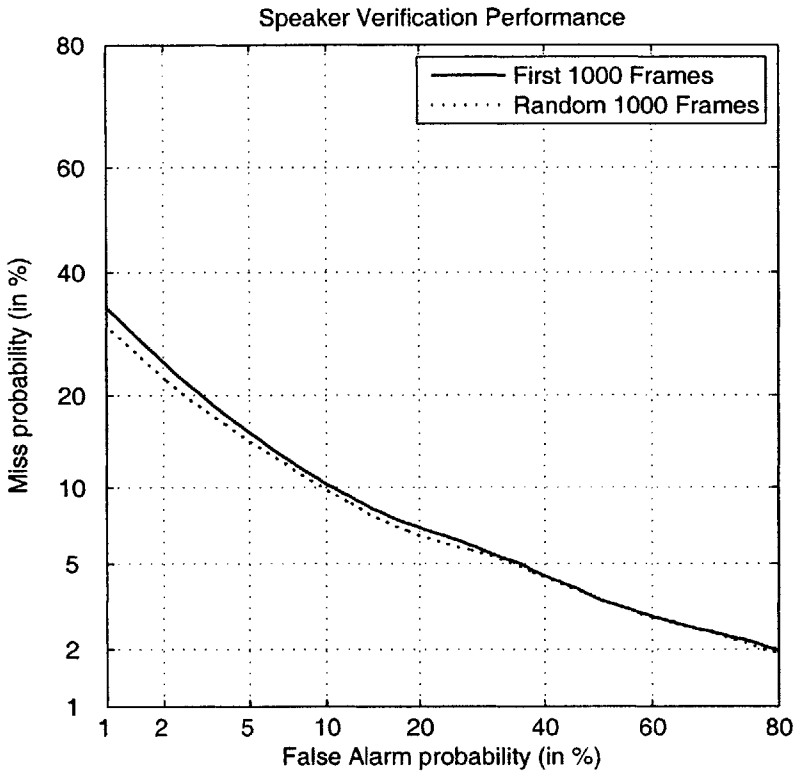

Figure 4-3: DET Curves for Two Training Image Selection Methods

The other 49 enrolled users are called in-set imposters. While they are not meant to be recognized by this particular support vector machine, data from their other recording sessions were used to train the support vector machine. The final group is called the out-of-set imposters and is made up of the 50 imposter subjects that only had one recording session each. The out-of-set imposters are completely unknown to the support vector machines used in the verification system. No images of these 50 subjects were used in the training step.

Because images of the in-set imposters were used to train the support vector machines, it should be easier for the verification system to distinguish between these users and the one correct user. We wanted to examine just how large a bias was created by testing the verification system with in-set and out-of-set imposters as opposed to just testing with out-of-set imposters.

We trained a set of support vector machines using 1000 face images from the office setting of the training data. We then tested the system using two sets of testing data. The first set contained all of the images from the office setting for all 100 subjects. The second testing set only tested images from the imposters and the one enrolled user that each support vector machine was supposed to recognize. The detection error tradeoff curves for these two trials is shown in Figure 4-4.

From the detection error tradeoff curve, it is clear that there is indeed some bias introduced by testing with both in-set and out-of-set imposters. Testing with both in-set and out-of-set imposters yielded an equal error rate of 10.18%. When we tested the system using only the out-of-set imposters the equal error rate was 10.67%. The small difference between these two equal error rates would indicate that any bias from testing using in-set imposters is not likely to have any major effect on our other experiments.

4.6 Individual Features

The last video-only experiment that we performed was to test the effectiveness of various facial features for verifying speakers. We experimented with four different

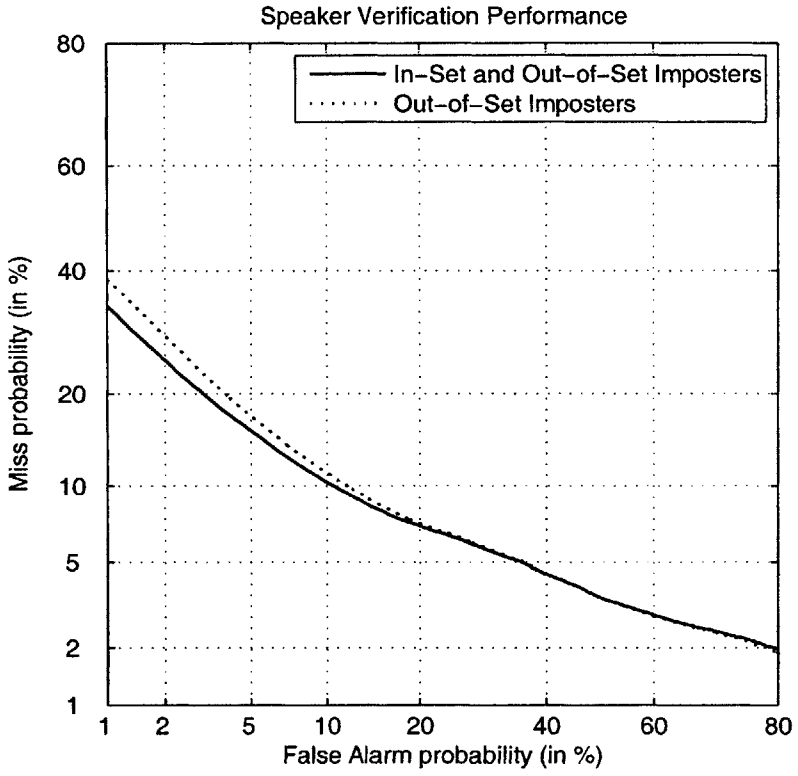

Figure 4-4: DET Curves for Testing With and Without In-Set Imposters

types of images, which are extracted from full frame images like the one in Figure 4-5. The first trial used loosely bounded face images. These images include the entire face and a small amount of the background environment in each corner of the image, as shown in Figure 4-6. The second image we used was a tightly bounded face image that included the majority of the face, but cropped any background as well as the chin and the top of the forehead. An example of the tightly bounded face image is in Figure 4-7. Third, we tested the system using mouth images, which include only the lips and a small amount of the surrounding area, as in Figure 4-8. Finally, we used images that include the part of the face from just under the nose to the top of the eyebrow. An example of this type of image can be seen in Figure 4-9.

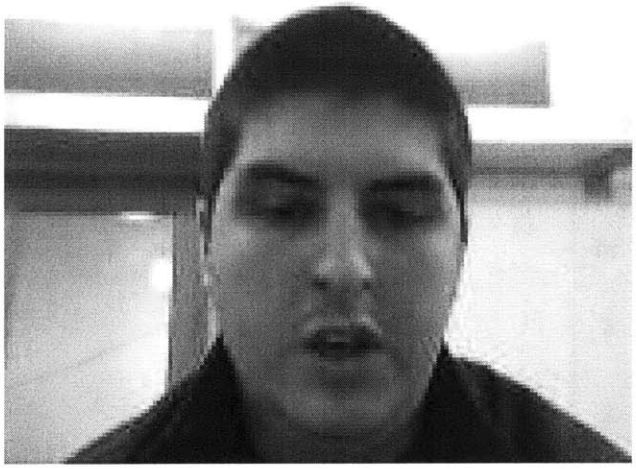

Figure 4-5: Sample Full Frame Image

We trained four verification systems, one for each type of image. The training sets include 1000 images from the office environment for each subject. The testing set included all the images from the office environment. The detection error tradeoff curves for each image type are plotted in Figure 4-10.

The tightly-bounded face images performed the best, producing an equal error rate of 10.18%. The mouth images had the second best score with 13.04%, followed by the loosely-bounded face images with an equal error rate of 14.85%. Finally, the nose-to-brow images returned an equal error rate of 17.51%.

Figure 4-6: Sample Loosely-bounded Face Image

Figure 4-7: Sample Tightly-bounded Face Image

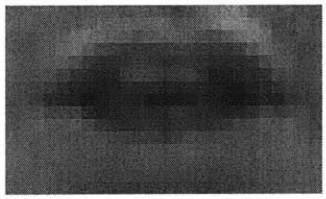

Figure 4-8: Sample Mouth/Lip Image

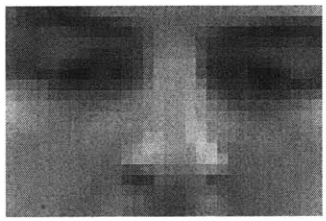

Figure 4-9: Sample Nose-to-Eyebrow Image

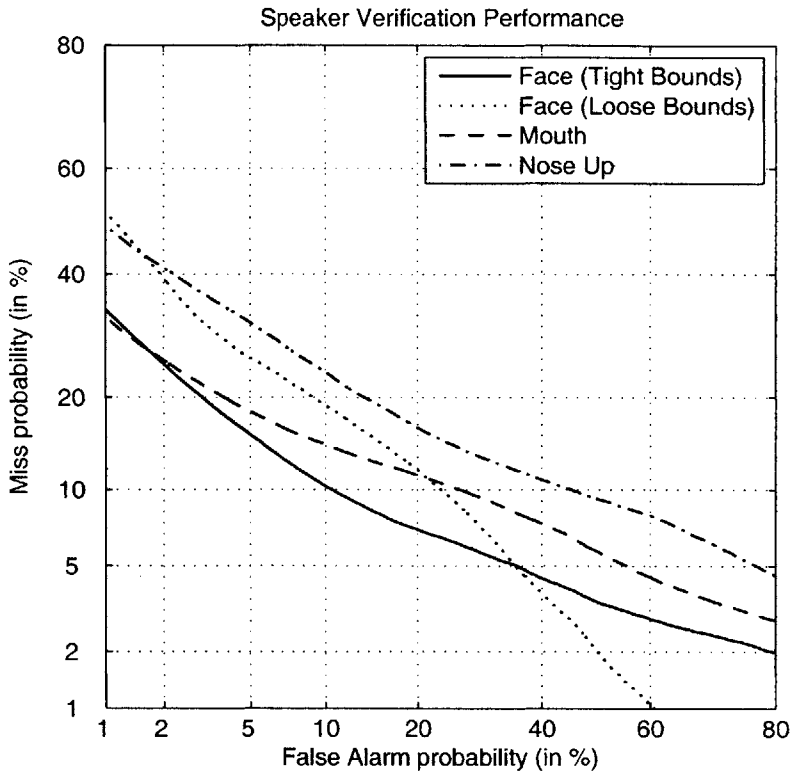

Figure 4-10: DET Curves for Four Types of Face/Face Feature Images

It is interesting to see how much of an effect a little bit of background in the image can have on the equal error rate of the verification system. The little bit of background in each corner of the image, along the small part of the face that is visible in the loosely-bounded face images but not the tightly-bounded versions, was able to increase the equal error rate by nearly five percent points.

Another surprising discovery is that the mouth images were able to outperform the nose-to-brow images. The mouth images contain the part of the face that moves the most while a subject is speaking. The nose-to-brow images include the part of the face that moves very little, aside from blinking. Yet the mouth images yielded an equal error rate nearly 4.5 percentage points lower than the nose-to-brow images.

Chapter 5

Audio Experiments

While the work for this thesis was mostly focused on video and multi-modal person verification systems, we could not study the multi-modal case without an independent audio speaker verification system. We trained the audio verification system as described in Section 5.1 and present the results of the system in Section 5.2. All the audio was extracted from the audio-visual recordings and stored in uncompressed WAV file format. All of the audio is 16kHz, 16 bit, mono sound.

5.1 Training the Audio Verification System

The audio speaker verification system was developed by the Spoken Language Systems Group at MIT. The system uses an automatic speech recognition (ASR) dependent method of speaker verification based on work described in [11] and [12].

The verification system actually consists of two main parts. One part handles speech recognition, while the other produces the verification scores. The first step in building the verification system is to train a model for each enrolled speaker. For each input utterance from the training data set, the speaker verification system produces a feature vector. The dimensionality of these feature vectors is then reduced using principal component analysis. The reduced feature vectors can then be used to train a model for each individual speaker.

To test our speaker verification system, we first ran each test utterance through

a speech recognizer. We used the SUMMIT speech recognizer, which is described in [6]. SUMMIT is a segment-based recognizer that uses both landmark and segment classifiers to produce the best hypothesis for phonetic segmentation. This hypothesis will be used by the verification part of the system to produce the score for each test utterance. Using a speech recognizer as part of the verification process will prevent playback attacks. If the system were fully text-independent, meaning the user could speak any word or phrase during verification, the audio from one successful attempt could be replayed in the future to gain access to the system. When using a recognizer, the user is given a phrase to speak and if the recognizer determines that the user spoke a different phrase, the user is likely to be rejected.

Independent from the recognition step, a reduced feature vector was produced for each test utterance in the same manner as the feature vectors for the training utterances. The reduced feature vector was used along with the phonetic segmentation hypothesis for comparison to the trained individual speaker models. A final verification score was then produced for each speaker model. The higher the score for a particular speaker, the more closely the utterance matched that speaker's model. Once scores are produced for all the test utterances, the scores can be used to determine the equal error rate of the system.

Full details for the speaker verification system can be found in [11] and [12].

5.2 Matched, Mismatched, and Mixed Training Sets

The audio verification system was run with each of the 12 possible combinations of training and testing environments, just like the video system in Section 4.3. The equal error rates for each of the training and testing pairs can be seen in Table 5.1. The matched condition equal error rates are 3.20% for the office setting, 12.28% in the cafe, and 13.48% for the street recordings. It is not surprising that the office setting has the best equal error rate and the street condition has the worst. The office setting typically had the least amount of background noise of the three environments, while the street typically had the most. These equal error rates are 7-10 percentage points

lower than the corresponding error rates for the video-only system. The relatively low quality of the video could be partially to blame for there being such a large discrepency between the audio and video verification system error rates.

The mismatched conditions produced equal error rates from 25.79% up to 41.42%. These equal error rates are worse than the video scores in every case, sometimes by more than 21 percentage points. So, while the matched condition audio scores were significantly lower than the matched video scores, the mismatched audio scores are significantly worse for each corresponding pair.

Finally, the mixed training condition scores are once again the best for each testing condition and with equal error rates of 2.45%, 5.80%, and 8.34% for office, cafe, and street testing, respectively, are lower than the video-only equal error rates for each testing condition. However, whereas the mixed-environment training for the video verification system used approximately the same amount of training data as the single-environment cases, the mixed-environment testing for the audio verification system used three times as much data as the rest of the experiments. In the mixed-environment experiments, the system was trained with the full audio samples from all 27 digits recordings from each of the enrolled users. It is likely that this is part of the reason for the mixed-environment equal error rates being lower than the single-location results.

		Testing Environment		
		Office	Cafe	Street
Training Environment	Office	3.20 %	30.13 %	38.24 %
	Cafe	37.71 %	12.28 %	25.79 %
	Street	41.42 %	28.26 %	13.48 %
	Mixed	2.45 %	5.80 %	8.34 %

Table 5.1: Equal Error Rates for Audio-Only Training/Testing Environment Pairs

Chapter 6

Multi-modal Experiments

After performing several experiments on the video data and audio data independently, it was time to combine the video and the audio to create a multi-modal speaker verification system. Audio was extracted from the original recording and used to train an audio speaker verification system. The scores returned by this system were combined with scores from the video speaker verification system to create a multi-modal speaker verification system that would hopefully be more accurate then either the audio or the video system could be on its own.

We combined the scores from each single-mode system using a weighted average. We attempted to find the best weights for the audio and video scores in Section 6.1. Choosing the best weight found, we ran experiments to compare the equal error rates of either single-mode system with the multi-modal system. Finally, we tested the equal error rates for the multi-modal system with both matched and mismatched environments as well as mixed-environment testing.

6.1 Weighted Average of Audio and Video Scores

After training and testing of both the audio and video speaker verification systems, we combined the scores by using a weighted average of the two scores. Unlike the video system which gave a score for each frame of a video, the audio system produced a frame average for each audio recording. Because the audio system gave only one score

per video, we averaged the frame scores for each video and combined that average video score with the audio scores.

For this experiment, we used audio and video that was recorded in the office setting. The video verification system used 1000 tightly-bounded face images per subject for training. Once the average score for each video was produced, it was combined with the audio score for that same video. When we calculated the weighted average, the video weight and audio weight added up to one. We began with an audio weight of zero and video weight of one. We then increased the audio weight by .05 and decreased the video weight by the same amount. We computed the equal error rate for each pair of weights to approximate the optimal weights. The equal error rates for each pair of weights are shown in Figure 6-1.

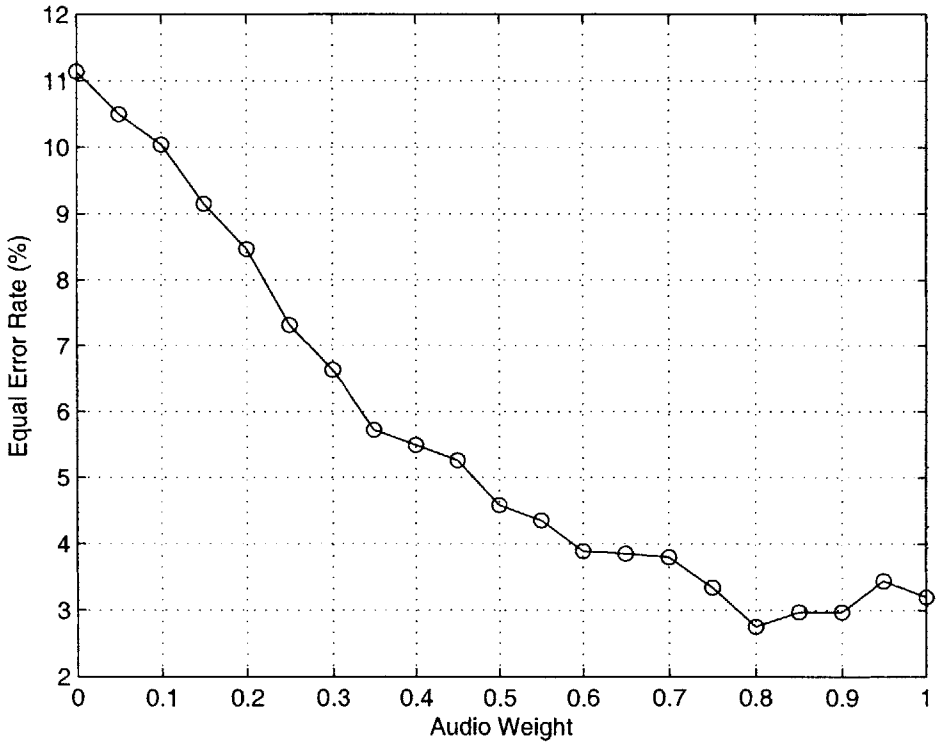

Figure 6-1: Equal Error Rates for Varying Audio Weights

For the office setting, the optimal weights were 0.80 for the audio and 0.20 for the video. These weights make sense as the equal error rate for the audio was lower than the equal error rate for the video. Therefore the audio would be more useful for verification, at least in the office setting. We examine the weights and equal error rates for other settings in Section 6.3.

6.2 Audio Versus Video Versus Multi-Modal

After determining the best weights for combining the audio and video scores. We wanted to compare the equal error rates for the single-mode systems with the equal error rate of the multi-modal system. The detection error tradeoff curves for each of the three systems is plotted in Figure 6-2. All three systems were trained and tested in the office setting. The single-mode video system is clearly the worst for speaker verification for the office setting, with an equal error rate of 10.18%. The audio system is much better, with an equal error rate of 3.20%. However, the multi-modal system is able to improve upon both of these systems, yielding an equal error rate of 2.75%. While this is a small absolute improvement in equal error rate over the video, it is almost a 15% relative improvement, which is certainly significant. Even with the low equal error rate of the single-mode audio system, it is promising that combining audio with video scores was still able to produce a more accurate system, despite the fact that the video had over three times the equal error rate of the audio system.

6.3 Matched, Mismatched, and Mixed Training Sets

After seeing the results from the audio system, we wanted to see what kind of improvement in equal error rate we could get in other environments by combining audio and video scores into a multi-modal system. We trained audio and video systems with data from each of the three environments and also with mixed-environment data. We then tested each of the systems separately with data from the three environments. For each environment, we tried every pair of weights with a 0.05 granularity, as de-

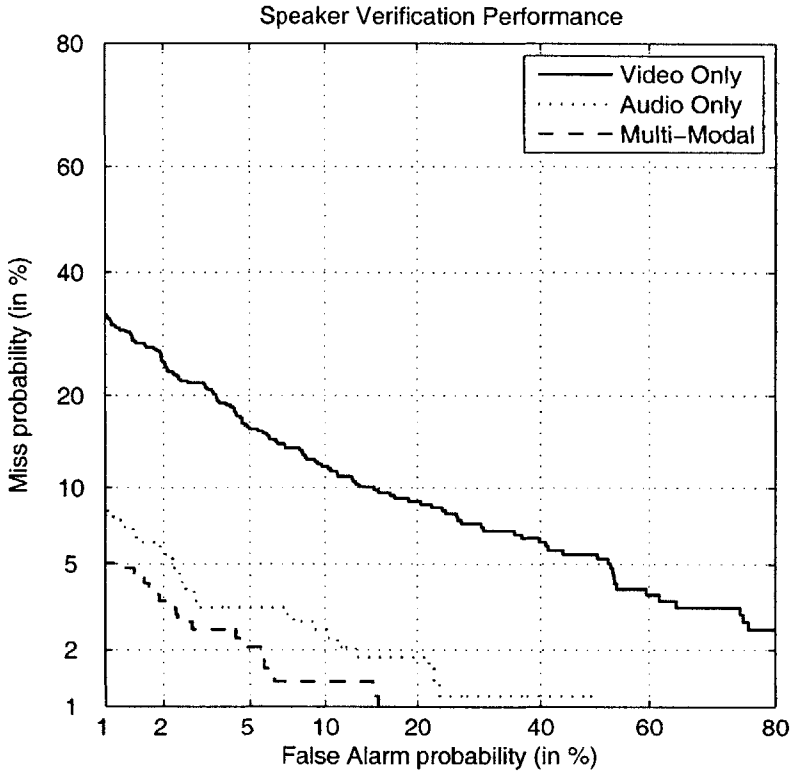

Figure 6-2: DET Curves for Audio, Video, and Multi-Modal Verification Systems Trained and Tested in the Office Environment

scribed in Section 6.1. The optimal weights were determined for each setting and the equal error rates were calculated.

Table 6.1 shows the best weights for averaging audio and video scores varied greatly in different environments. There were three pairs of training and testing locations whose best equal error rates came from basing 80% of the multi-modal scores on the audio verification system and only 20% on the video system. There are also three location pairs that used the audio and video scores almost equally. Finally, when training in the office setting and testing in the street environment, the best equal error rate resulted from basing 65% of the multi-modal score on the video and only 35% on audio.

	Testing Environment			
		Office	Cafe	Street
Training Environment	Office	0.80	0.70	0.35
	Cafe	0.55	0.75	0.80
	Street	0.45	0.75	0.75
	Mixed	0.80	0.70	0.55

Table 6.1: Optimal Audio Weights for Training/Testing Environment Pairs

The best equal error rate for each pair of training and testing sets found using the multi-modal verification system can be found in Table 6.2. When the multi-modal system uses the best weights for each environment, we call this the oracle system, because a real world verification system would not know which weights would produce the best scores for a particular test until after testing had been conducted. The error rates for matched environments are 2.75% for the office setting, 9.60% for the cafe setting, and 10.30% for data collected in the street environment. It is not surprising that the error rate increases from office to cafe to street environments. The level of audio noise and amount of lighting variety in the data increases from office to cafe to street environments making the verification task easiest for the office and most difficult for the street. We saw the same increases in error rates for the mixed-environment training experiments. The office testing error rate was the lowest with 1.60% and the street the highest with 4.02%.

Fortunately, no matter what the best weights for audio and video scores happen to

Location		System			
Training	Testing	Video	Audio	Multi-Modal (Oracle)	Multi-Modal (70% Audio)
Office	Office	11.14 %	3.20 %	2.75 %	3.80 %
Office	Cafe	22.76 %	30.13 %	16.29 %	16.29 %
Office	Street	15.38 %	38.24 %	14.03 %	15.61 %
Cafe	Office	25.17 %	37.71 %	23.93 %	24.03 %
Cafe	Cafe	21.20 %	12.28 %	9.60 %	10.03 %
Cafe	Street	22.39 %	25.79 %	18.71 %	19.14 %
Street	Office	26.77 %	41.42 %	25.39 %	25.52 %
Street	Cafe	24.66 %	28.26 %	18.53 %	18.97 %
Street	Street	20.35 %	13.48 %	10.30 %	10.63 %
Mixed	Office	11.21 %	2.45 %	1.60 %	2.29 %
Mixed	Cafe	12.95 %	5.80 %	3.75 %	3.75 %
Mixed	Street	10.14 %	8.34 %	4.02 %	5.12 %

Table 6.2: Equal Error Rates for Multi-Modal Training/Testing Environment Pairs

be and no matter what environments are used for training and testing, if the proper weight can be found, the multi-modal verification system always out-performs either of the single-mode systems. On average, the relative improvement of the multi-modal oracle system over using video alone is 40.4%. The average relative improvement over using audio alone is 35.8%. If the multi-modal equal error rate is compared to the best of the two single-mode systems, the average relative improvement is 22.49%.

6.4 Single-Weight Multi-Modal Verification

While the results in Section 6.3 are promising, it is important to note that all of the scores were achieved using the optimum weights for combining the audio and video scores. When the same pair of weights was used for each of the 12 training and testing environments combinations, the results show much smaller improvements. The best overall improvement was achieved using a 70% contribution from audio and a 30% contribution from video. The equal error rates for each training and testing environment can be seen in Table 6.2. The average improvement for the 12 possible cases was 8.51% against the video-only system and 28.30% against the audio-only system. Finally, if you compare the multi-modal system using the single set of weights

to the best of the two single-mode systems, the average improvement is only 4.74%. Clearly, the proper choice of weights for combining audio and video verification scores is crucial if the multi-modal system is to offer a significant benefit over a single-mode system.

Chapter 7

Conclusions

7.1 Summary

This thesis examined the problem of multi-modal speaker verification using limited enrollment data recorded in low resolution using a web-camera. Multiple factors for improving the single-mode video-only person verification system were tested first. Then audio and video verification scores were combined to test the capabilities of multi-modal speaker verification.

7.1.1 Video-Only Speaker Verification Results

In Chapter 4, we examined the effects of several aspects of video-only speaker verification. We found that the equal error rate for performing verification on single frames is nearly identical to the equal error rate produced when using simple methods for combining frame scores into a single video score. We also examined the effect of training set size and found that using more training images leads to lower error rates, but that as training sets become large the computational cost outweighs any gain that increasing training set size can have. Section 4.5 showed that including in-set imposters when testing our system did not introduce a large bias.

Next, we experimented with different combinations of training and testing environments. These experiments demonstrate that if training and testing cannot both

be performed in a controlled environment, the equal error rate can be kept at a reasonable level by diversifying the training environment. Testing in the cafe and street locations produces scores similar to those for the office environment by training with mixed environment data.

Finally, we explored the effectiveness of using several types of face and facial feature images. We found that the best single image type was a tightly bounded face image.

7.1.2 Multi-Modal Speaker Verification Results

Chapter 6 explored the possibilities of combining scores from single-mode audio and video verification systems to attempt to improve upon the equal error rates of either system taken independently. We found that the multi-modal system can consistently outperform either single-mode system, however, this is only possible if the proper weights can be found for combining single-mode scores. Using the optimum weights for each training and testing environment combination produced equal error rates that were over 22% lower on average than the best of the two single-mode systems. However, this improvement is less than 5% when using a single set of weights for all environments. Using optimum weights and mixed environment training data, the multi-modal system was able to achieve equal error rates below 5% and in the case of the office testing environment the equal error rate was 1.60%, which is the best equal error rate for any experiment we performed.

7.2 Future Work

In exploring the possibilities of our audio-visual person verification system, there were a number of avenues that we did not explore, but which might prove useful in improving the robustness and accuracy of our system. In future work we hope to address these items. One aspect of our system that deserves further examination is the method by which training images are selected. Using randomly selected images from each subject proved to be a better system than simply starting from the beginning

and choosing consecutive images, randomly selecting images does not address the fact that there are some images that may be of little use to the verification system. For instance, an image could have particularly bad lighting or the coordinates for the face location in the frame could have been incorrect. Training with such images will only serve to confuse the verification system. If some measure could be developed to discriminate between good and bad training frames, this could drastically improve the quality of the video side of the verification system.

Another area for exploration would be the effect of combining individual feature images extracted from one frame into a single multiple-feature image and using these images for training and testing. Extracting features separately allows for lighting to be normalized for each feature independently which can lead to less environmentally-caused lighting variation. An example of where this might be useful is if one side of the user's face is in shadow and the other side is not. Being able to normalize the light from the right eye separate from the left eye could reduce the effect of the shadows on the face.

Finally, we would like to build a classifier to determine the best weights for combining scores from the audio verification system with scores from multiple individual facial features. This would provide a great way to determine which features are the most effective for the verification task.

Bibliography

[1] OpenCV. `http://sourceforge.net/projects/opencvlibrary/`.

[2] SvmFu. `http://fpn.mit.edu/SvmFu/`.

[3] XM2VTS. `http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/`.

[4] S. Ben-Yacoub, J. Luttin, K. Jonsson, J. Matas, and J. Kittler. Audio-visual person verification. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1580–1585, Fort Collins, Colorado, June 1999.

[5] N. Fox and R.B. Reilly. Audio-visual speaker identification based on the use of dynamic audio and visual features. In *Proc. of 4th Int. Conf. on Audio- and Video-Based Biometric Person Authentication*, pages 743–751, Guildford, UK, June 2003.

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

[7] T. Hazen, E. Weinstein, R. Kabir, and A. Park. Multi-modal face and speaker identification on a handheld device. In *Proc. of Workshop on Multimodal User Authentication*, pages 113–120, Santa Barbara, California, December 2003.

[8] T. Hazen, E. Weinstein, and A. Park. Towards robust person recognition on handheld devices using face and speaker identification techniques. In *Proc. of Int. Conf. on Multimodal Interfaces*, pages 19–41, Vancouver, Canada, November 2003.

[9] B. Heisele, P. Ho, and T. Poggio. Face recognition with support vector machines: Global versus component-based approach. In *Proc. of IEEE Int. Conf. on Computer Vision*, volume 2, pages 688–694, Vancouver, Canada, July 2001.

[10] B. Maison, C. Neti, , and A. Senior. Audio-visual speaker recognition for video broadcast news: some fusion techniques. In *Proc. of IEEE Workshop on Multimedia Signal Processing*, Copenhagen, Denmark, September 1999.

[11] A. Park and T. Hazen. ASR dependent techniques for speaker identification. In *Proc. of Int. Conf. on Spoken Language Processing*, pages 1337–1340, Denver, Colorado, September 2002.

[12] A. Park and T. Hazen. A comparison of normalization and training approaches for ASR-dependent speaker identification. In *Proc. of Int. Conf. on Spoken Language Processing*, pages 2601–2604, Jeju Island, Korea, October 2004.

[13] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, Glass J, , and T. Darrell. Visual speech recognition with loosely synchronized feature streams. In *Proc. of IEEE Int. Conf. on Computer Vision*, pages 1424–1431, 2005. `http://doi.ieeecomputersociety.org/10.1109/ICCV.2005.251`.

[14] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 511–518, Kauai, Hawaii, December 2001.