Exploration of Small Enrollment Speaker Verification on Handheld Devices

by

Ram H. Woo

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

Bachelor of Science in Electrical Engineering and Computer Science

and

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2005

© Massachusetts Institute of Technology 2005. All rights reserved.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
JUL 18 2005
LIBRARIES

Author .....
Department of Electrical Engineering and Computer Science
May 18, 2005

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

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

ARCHIVES

Exploration of Small Enrollment Speaker Verification on Handheld Devices

by

Ram H. Woo

Submitted to the Department of Electrical Engineering and Computer Science
on May 18, 2005, in partial fulfillment of the
requirements for the degrees of
Bachelor of Science in Electrical Engineering and Computer Science
and
Master of Engineering in Electrical Engineering and Computer Science

Abstract

This thesis explores the problem of robust speaker verification for handheld devices under the context of extremely limited training data. Although speaker verification technology is an area of great promise for security applications, the implementation of such a system on handheld devices presents its own unique challenges arising from the highly mobile nature of the devices. This work first independently analyzes the impact of a number of key factors, such as speech features, basic modeling techniques, as well as highly variable environmental/microphone conditions on speaker verification accuracy. We then present and evaluate methods for improving speaker verification robustness. In particular, we focus on normalization techniques, such as handset normalization (H-norm), zero normalization (Z-norm) as well as model training methodologies (multistyle training) to minimize the detrimental impact of highly variable environment and microphone conditions on speaker verification robustness.

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

Acknowledgments

I would first like to express my gratitude to my thesis advisor T.J. Hazen for his kind patience and guidance. His invaluable mentorship throughout this past year has helped me to grow as a researcher. Furthermore, his insightful comments have been critical in helping me navigate through this project.

I would also like to acknowledge the support of the members of the Spoken Language Systems Group as well as express my appreciation for welcoming me into the group. Specifically, I would like to thank Alex Park for his help in numerous brainstorming and debugging sessions. I also appreciate his willingness to field my many questions.

Finally, I would like to thank my parents, sister, Namiko, as well as my dear friends for their continual love, support, and encouragement. Without them I would be lost.

This research was made possible by the support of Intel Corporation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As technological improvements allow for the development of more powerful and ubiquitous handheld devices, such as PDAs and handheld computers, there also exists a need for greater security. No longer merely novelty items, handhelds have advanced far beyond the realm of simple calendars and now have the ability to perform a myriad of computationally complex tasks. Hence, reliable ways to control access to sensitive information stored on these devices must be devised.

Currently, the most prevalent security mechanism is the text-inputted password. Although simple in implementation, this system is hobbled by a number of handicaps. Its effectiveness is highly dependent on the use of hard-to-remember string / digit combinations which must be frequently changed. However, in practice, users opt for simple pass-phrases which are rarely, if ever, altered providing little actual security. Furthermore, the small keyboard layouts often found on handheld devices make the task of frequently inputting passwords a tedious affair. Ultimately, the text-inputted password does not protect the user in situations where both the device and password are stolen.

One viable alternative, which promises greater flexibility and ease of use, is the integration of speaker verification technology for secure user logins. Speaker verification provides an additional biometric layer of security to protect the user. The focus

of this work is to investigate the use and effectiveness of speaker verification for use on small, handheld devices.

1.1 Motivation

Driven in part by its promising potential in security applications, speaker verification has been a heavily researched field. Although the term speaker *verification* is oftentimes used interchangeably with speaker *identification*, these terms refer to distinct, albeit, closely related tasks. The goal of identification is to determine, given a sample of speech, a speaker's identity from a cohort of previously enrolled users. Verification, however, takes both a speech utterance as well as the purported user's identity and verifies the authenticity of the claim.

Previous work on speaker verification systems can be largely sub-divided into two major domains: telephone-based and vestibule security. Telephone-based verification systems have a number of applications, particularly in transactions requiring secure access to financial information (i.e. telephone-shopping, bank account balance, etc). Examples include work by Boves et. al. [4] and Lamel and Gauvain [8]. Commercial systems developed by Scansoft and Nuance are tailored towards industry areas, such as the healthcare and financial sectors, that require high levels of security to protect sensitive customer account information [1], [2]. In addition to providing security, commercial speaker verification systems allow companies to reduce costs by replacing expensive live call centers with automated systems for speaker verification.

Vestibule security, the second major domain of speaker verification, is frequently portrayed in Hollywood movies and focuses on the fast and secure physical access to restricted locations. Speaker verification allows for contact-less activation and eliminates the risks of stolen or lost keys / passwords / keycards inherent to key-based entry mechanisms. Examples include work by Morin and Junqua [11] as well as Doddington [5]. Furthermore, speaker verification can be used in conjunction with

various other modalities (fingerprint, keypad, and/or face verification) to maximize flexibility, convenience, and performance in vestibule security.

1.2 Technical Challenges

Although speaker verification technology is an area of great promise for security applications, the implementation of such a system on handheld devices presents its own unique challenges.

1.2.1 Environmental Conditions

One of the largest challenges in implementing speaker verification on handheld devices arises from the handheld's greatest attribute: mobility. Unlike vestibule security systems, handheld devices experience use in highly variable acoustic environments. Through the course of just one day, a user may activate their handheld device to transfer data files at the office, check e-mail while eating in the cafeteria, and play audio files as they are crossing a busy street intersection. In each environment, variations in the acoustical conditions will alter the sound of a user's speech leading to intra-speaker variability [7]. This *intra*-speaker variability complicates the task of differentiating speakers based on *inter*-speaker variations leading to reduced accuracy in speaker verification.

Additionally, speaker verification systems must also be robust against varying degrees of background noise inherent to each environment. Although environments such as a quiet office, with little ambient noise, are ideal when conducting speaker verification, it is impossible (as well as highly undesirable) to constrain users to such locations before granting access to the handheld device. Thus, the issue of minimizing performance degradation, due to distortions introduced by wind, rain, background speakers, road traffic, etc., is critical in the development of speaker verification for use on handhelds.

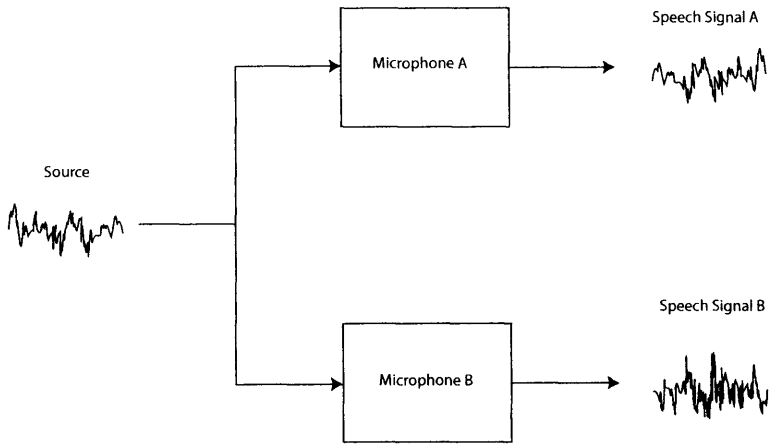

Figure 1-1: Block Diagram Illustrating Effect of Microphones on Speech Signal

1.2.2 Microphone Variability

The wide variability in the microphones used with handheld devices can also have a substantial impact on performance in speaker verification systems. Microphones introduce channel effects which are both linear and non-linear in nature and can be difficult or impossible to remove. These channel effects create distortions in a user's speech signal as illustrated in Figure 1-1. Furthermore, different microphones can have highly dissimilar transfer functions. Hence, speaker verification systems experience large degradations in accuracy when different microphones are used for enrollment and verification.

1.2.3 Low Enrollment Data

The problem of maintaining robustness and verification accuracy on handheld devices is further compounded by the limited enrollment data upon which to develop a handheld based speaker verification system. Unlike test systems developed for use in feasibility studies, real world systems are constrained by usability issues. One of

the foremost concerns is ease of use. Handheld based verification systems must allow for the quick and easy enrollment of new users. However, this ease of use comes at a cost. Short enrollment sessions provide limited data, preventing the training of robust, phonetically balanced speaker models. Low enrollment data also exacerbates the deleterious effects of environmental conditions and microphone variability.

1.3 Goals

While the task of developing a speech based speaker verification system has been a topic of substantial research, much of the work has centered around scenarios where a large and phonetically rich corpus of training data is available. This thesis departs from that theme to explore the problem of robust speaker verification for handheld devices under the context of extremely limited training data. This work first analyzes the impact of a number of key factors, such as speech features, environmental conditions, training methodologies, and normalization techniques, on speaker verification performance independently. These factors are then examined in conjunction in order to identify how best to maximize a system's overall robustness and accuracy.

1.4 Outline

The rest of this work is organized as follows:

- Chapter 2 provides an overview of the basic techniques of speaker verification.
- Chapter 3 describes in detail the process of data collection.
- Chapter 4 explores various experiments in speaker verification. The chapter begins by discussing basic speaker verification modeling, analyzing the impact of speech features such as MFCCs as well as differing speech models (i.e. boundary, segment, and frame based modeling). The effects of mismatched conditions on

speaker verification performance are then explored. In particular microphone, environment, and vocabulary effects are analyzed. In order to improve verification robustness, we also investigate methods for multistyle testing as well as the H-norm and Z-norm normalization techniques. Finally, we discuss the impact of knowledge on system performance.

- Chapter 5 summarizes and draws together concluding remarks on the paper.

Chapter 2

Basic Techniques of Speaker Verification

2.1 Background

Given a speech segment from an alleged user, the goal of speaker verification is to either correctly authenticate the speaker's identity or to reject the speaker as an imposter. The implementation of a speaker verification system consists of a two step procedure:

- **Enrollment:** Process by which speaker models are trained for the system users. Each user engages in an enrollment session in which speech data is collected from the user and is utilized to train a speaker model for the specific user. This is analogous to designing a biometric lock with the speaker's voice as the key.

- **Verification:** Testing phase of the system. A purported user attempts to log onto the system, as a previously enrolled user, by reciting an utterance. This new speech sample is then tested against the enrolled user's speaker model and a score is computed. The final decision of "accept" or "reject" is determined through a comparison of the speaker's score against a predetermined threshold.

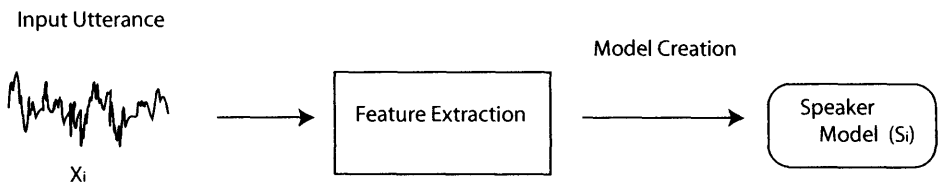

Figure 2-1: Block Diagram of Enrollment Procedure

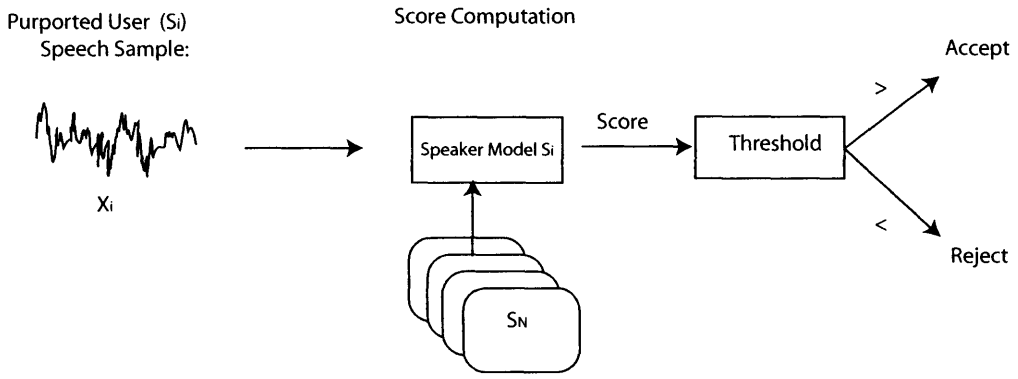

Figure 2-2: Block Diagram of Verification Procedure

A speaker with a score greater than the threshold is accepted while a speaker with a score lower than the threshold is rejected.

Additionally, speaker verification techniques can be categorized into two major classifications: **Text - Independent** and **Text - Dependent**.

2.1.1 Text-Independent Speaker Verification

Under text-independent speaker verification, rather than prompting the user to pronounce a certain set of phrases, the vocabulary of the speaker is left completely

unconstrained. Traditionally, text-independent speaker verification techniques have largely been centered around the use of Gaussian Mixture Models (GMMs) [15]. Speaker models based on Mel-frequency cepstral coefficients (MFCCs), features that model the spectral energy distribution of speech, are trained using all utterances for a particular speaker. The GMM density model is characterized by a weighted linear combination of Gaussian densities, each parameterized by a mean and variance. Given a N-dimensional feature vector $\mathbf{x}$, the gaussian mixture density, for a given speaker S, is characterized as:

$$p(\mathbf{x}|S) = \sum_{i=1}^{M} w_i p_i(\mathbf{x}) \tag{2.1}$$

Where $\sum_i w_i = 1$. Each $p_i(\mathbf{x})$ is defined as a multivariate Gaussian density with covariance matrix $\sum_i$, and mean u_i.

$$p_i(\mathbf{x}) = \frac{1}{(2pi)^{N/2}|\sum_i|^{1/2}} exp\{-\frac{1}{2}(\mathbf{x}-\mathbf{u_i})'(\sum_i)^{-1}(\mathbf{x}-\mathbf{u_i})\} \tag{2.2}$$

2.1.2 Text-Dependent Speaker Verification

A competing, text-dependent, approach to the task of speaker verification is the MIT CSAIL Speaker Adaptive ASR-based system [12]. Text-dependent speaker verification constrains the speaker to a limited vocabulary and directs the user to speak fixed phrases. Text-dependent verification systems differ from GMM based systems by taking into account phonetic knowledge when developing speaker models. This allows the system to utilize differences in phonetics events when making determinations. As described in [13], let $\mathbf{X}$ represent the set of feature vectors, $\{x_1, ..., x_n\}$, extracted from a particular spoken utterance while S($\mathbf{X}$) will denote the speaker of the said utterance. During training, speaker-dependent phone dependent (SD-PD)models are created from phonetically transcribed enrollment data. Hence, each speaker, S, is represented by a set of models, $p(x|S, \phi)$, where ϕ represents a phonetic unit and

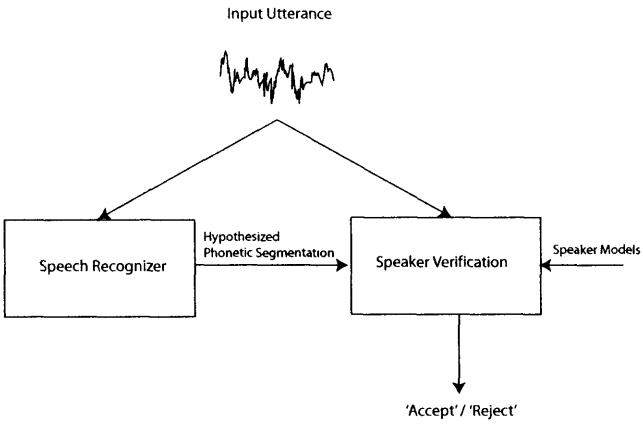

Figure 2-3: Block Diagram of Overall Speaker Verification System

$\phi(x_k)$ denotes the underlying phonetic unit of the feature vector x_k.

In the testing phase, a phonetic transcription is generated for each test utterance and then used to score the segment with a speaker-dependent phonetic model. Therefore, the most likely phonetic unit $\hat{\phi}(x_k)$ is assigned to each feature vector x_k and a speaker-dependent phone-dependent conditional probability, $p(x|S, \hat{\phi}(x))$ is computed. A hypothesized speaker, $\hat{S}(\mathbf{X})$, is then determined as follows:

$$p(\mathbf{X}|S, \hat{\Phi}(\mathbf{X})) = \prod_{\forall x} p(x|S, \hat{\phi}(x)) \tag{2.3}$$

$$p(\mathbf{X}|S, \hat{\Phi}(\mathbf{X})) > \theta \Longrightarrow accept \tag{2.4}$$

$$p(\mathbf{X}|S, \hat{\Phi}(\mathbf{X})) < \theta \Longrightarrow reject \tag{2.5}$$

Where θ is a threshold value.

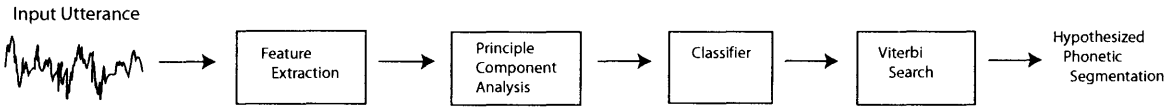

Figure 2-4: Block Diagram of Speech Recognition Component (SUMMIT)

2.2 Overview of Speaker Verification System

Developed by the Spoken Language Systems Group (SLS) at MIT, the speaker verification system used in the following experiments, is a parallel two-component process consisting of speech recognition and speaker verification. The system was developed under the SAPPHIRE research environment.

2.2.1 SUMMIT

The SLS SUMMIT speech recognizer, the first component of the speaker verification system is a segment-based recognizer which combines segment and landmark based classifiers. SUMMIT takes an inputted speech utterance and maps each acoustic observation to a hypothesized phonetic segmentation. Details of the SUMMIT system can be found in [6]. This hypothesis is then outputted for later use in the speaker verification component. This procedure is illustrated in Figure 2-4

2.2.2 Speaker Verification Module

The second major component of the overall system is the speaker verification module illustrated in Figure 2-5. For each input waveform, the verification module first conducts feature extraction

1. **Frame-Based Observations:** regular time intervals (i.e. 10 ms)
2. **Segment-Based Observations:** variable-length phonetic segments
3. **Landmark-Based Observations:** regions surrounding proposed phonetic boundaries

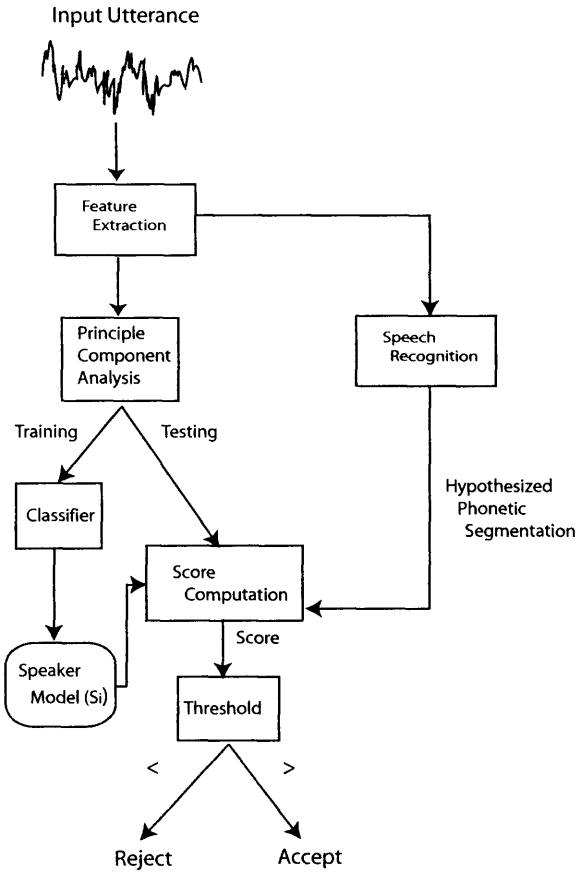

Figure 2-5: Block Diagram of Speaker Verification Component

Although the feature extraction is typically based on normalized Mel-frequency cepstral coefficients (MFCCs), other speaker-specific features such as pitch can be used[1]. Each observation (frame, segment, or landmark), is then represented by an M-dimensional feature vector $\mathbf{x_i}$ that is created by concatenating N different averages of the region surrounding the current observation. For example, if 8 (i.e. N=8) different 14 coefficient MFCC vectors are used, each feature vector $\mathbf{x_i}$ would be of size M=112.

Once the feature vectors are extracted, they then undergo principal component analysis (PCA) to reduce the dimensionality of the vectors to 50. PCA attempts to decorrelate the acoustic measurements by projecting the Gaussian data onto orthogonal axes that maximize the variance of each projection. Dimensionality reduction is then made possible by keeping only the components of maximal variance.

At this stage, the speaker verification module can progress along one of two divergent paths:

- **Training:** Under the training modality, individual speaker models are trained from the reduced feature vectors. Training is conducted under one of two procedures. For segment or landmark based features, speaker-dependent phone GMMs are first trained for each speaker. In our system, these phone specific GMMs are then collapsed and combined to create a speaker-specific global GMM. Our frame based models, however, are trained under a slightly different process whereby only speaker-specific global GMMs are created. Frame-based training bypasses the creation of phone specific models. Although the two training methodologies are procedurally different, the resulting GMM speaker models are analogous.

- **Testing:** Under the testing modality, the reduced feature vectors are fed into the speaker verification module. Additionally, the hypothesized phonetic segmentation determined from the speaker independent speech recognizer is also

[1]One thing to note is that the features used in the speaker verification module need not be the same features used for the speech recognition module.

inputted. These feature vectors are then scored against pre-trained claimant models. Speech samples that prove to be a good match to a speaker's model produce positive scores while negative scores represent poor matches.

Chapter 3

Data Collection

In this chapter, we describe the task of data collection. For the experiments, a prototype Morro Bay handheld device, donated by Intel Corporation, was utilized.

3.1 Overview

In order to simulate scenarios encountered by real-world speech verification systems, the collected speech data consisted of two unique sets: a set of "enrolled" users and a set of "imposters". For the "enrolled" set, speech data was collected from 48 users over the course of (2) twenty minute sessions that occurred on separate days. In the "imposter" set, approximately 50 new users participated in (1) twenty minute session.

3.2 Phrase Lists

Within each data collection session, the user recited a list of name and ice cream flavor phrases which were displayed on the hand-held device. An example phrase list can be found in Table 3.1. In developing the phrase lists, the main goal was to produce a phonetically balanced and varied speech corpus. 12 list sets were created for "enrolled" users (8 male list sets / 4 female list sets) while 7 lists were created

for "imposter" users (4 male lists / 3 female) lists. Each "enrolled" user's list set contained two phrase lists which were almost identical, differing only in the location of the ice cream flavor phrases on the lists. The first phrase list was read in the "enrolled" user's initial data collection session, while the second list phrase was used in the subsequent follow-up session.

3.3 Environmental / Acoustic Conditions

In order to capture the expected variability of environmental and acoustic conditions inherent with the use of a hand-held device both the environment and microphone conditions were varied during data collection. For each session, data was collected in three different locations (a quiet office, a noisy hallway, and a busy street intersection) as well as with two different microphones (the built-in microphone of the handheld device and an external earpiece headset) leading to 6 distinct test conditions. Users were directed to each of the 3 locations, however, once at the location the person was allowed to roam freely.

3.4 Statistics

In total, each session yielded 54 speech samples per user. This yielded 5,184 examples from "enrolled" users (2,592 per session) and 2,700 "imposter" examples from users not in the enrollment set. Within the "enrolled" set of 48 speakers, 22 were female while 26 were male. For the "imposter" set of 50 speakers, 17 were female while 23 were male.

Office/External	Hallway/External	Intersection/External
alex park	alex park	alex park
rocky road	chocolate fudge	mint chocolate chip
ken steele	ken steele	ken steele
rocky road	chocolate fudge	mint chocolate chip
thomas cronin	thomas cronin	thomas cronin
rocky road	chocolate fudge	mint chocolate chip
sai prasad	sai prasad	sai prasad
rocky road	chocolate fudge	mint chocolate chip
trenton young	trenton young	trenton young
Office/Internal	Hallway/Internal	Intersection/Internal
alex park	alex park	alex park
peppermint stick	pralines and cream	chunky monkey
ken steele	ken steele	ken steele
peppermint stick	pralines and cream	chunky monkey
thomas cronin	thomas cronin	thomas cronin
peppermint stick	pralines and cream	chunky monkey
sai prasad	sai prasad	sai prasad
peppermint stick	pralines and cream	chunky monkey
trenton young	trenton young	trenton young

Table 3.1: Example of Enrollment Phrase List

Chapter 4

Experimental Results

4.1 Basic Speaker Verification Modeling

In this section, experiments were conducted on basic speaker verification modeling techniques. These tests were designed to identify the optimal acoustic-phonetic representation of speaker specific information for the collected Morro Bay speech corpus.

4.1.1 Experimental Conditions

Our speaker verification system relied on a speech recognition alignment to provide temporal landmark locations for a particular speech waveform. Furthermore, we assumed the speech recognizer to provide the correct recognition of phrases and the corresponding phone labels. In real world applications, this assumption is acceptable in situations where the user always utters the same passphrase. As described in [6], landmarks signify locations in the speech signal where large acoustic differences indicate phonetic boundaries. In developing landmark-based models, feature vectors consisting of a collection of averages of Mel-frequency cepstral coefficients (from eight different regions) surrounding these landmarks were extracted.

In the following experiments, enrolled users uttered one ice cream flavor phrase 4 times within a single enrollment session. This enrollment session took place within the

office environment with the use of an external earpiece headset microphone. During testing, identical environment and microphone conditions were maintained and the verification accuracy of previously enrolled users reciting the same phrase (from the enrollment session) was compared to dedicated imposters also speaking the same phrase.

4.1.2 Global Gaussian Mixture Models vs. Speaker-Dependent Phone-Dependent Models

As previously discussed in Chapter 2, current speaker verification techniques generally capture speaker specific acoustic information using one of two methods: Gaussian mixture models (GMMs) or speaker-dependent phone-dependent (SD-PD) models. In order to empirically determine which models resulted in the best fit, we performed verification experiments using MIT CSAIL's ASR-Dependent System coupled with phone adaptive normalization. Mathematically, for a given speaker S and phonetic unit $\hat{\phi}(\mathbf{x})$, the speaker score is:

$$Y(\mathbf{X}, S) = \frac{1}{|X|} \sum log[\lambda_{S,\hat{\phi}(\mathbf{x})} \frac{p(\mathbf{x}|S, \hat{\phi}(\mathbf{x}))}{p(\mathbf{x}|\hat{\phi}(\mathbf{x}))} + (1 - \lambda_{S,\hat{\phi}(\mathbf{x})}) \frac{p(\mathbf{x}|S)}{p(\mathbf{x})}] \tag{4.1}$$

Where $\lambda_{S,\hat{\phi}}$, represents the interpolation factor given that $n_{S,\hat{\phi}(\mathbf{x})}$ is the number of times the phonetic event $\hat{\phi}(\mathbf{x})$ is observed and τ is a tuning parameter.

$$\lambda_{S,\hat{\phi}(\mathbf{x})} = \frac{n_{S,\hat{\phi}(\mathbf{x})}}{n_{S,\hat{\phi}(\mathbf{x})} + \tau} \tag{4.2}$$

Further details of the phone adaptive normalization technique can be found in [13]. By utilizing phone adaptive normalization, speaker-dependent phone-*dependent* models are interpolated with a speaker-dependent phone-*independent* model (i.e. a global GMM) for a particular speaker. As τ, and thereby the interpolation factor $\lambda_{S,\hat{\phi}(x)}$ is adjusted, phone dependent and phone independent speaker model probabilities are combined in varying ratios.

τ	**Interpretation**	**EER**
0	SD-PD Models	51.56%
5	SD-PD / GMM Combo	10.42%
∞	global GMM	10.94%

Table 4.1: Identification error rates in relation to τ moving from 0 to 5 to ∞

Table 4.1 shows the verification equal error rates as τ is varied from 0 to ∞. Figures 4-1 and 4-2 display the corresponding detection/error tradeoff (DET) curves.

As can be seen, a global GMM performed substantially better than SD-PD models which produced an EER of 51.56%, roughly equal to that of random chance. This was not highly unexpected, however, as sparse enrollment data prevented the training of robust models at the phone level. Furthermore, for a majority of the phones, no training tokens existed. While the global GMM also suffered from limited enrollment data, it proved more robust to this issue as all available data was used to train a single large model as opposed to multiple smaller refined models.

As shown in Figure 4-2, however, an absolute performance increase of 0.52% was achieved by combining phone dependent and GMM speaker model probabilities. This result mirrored previous experiments conducted in [13].

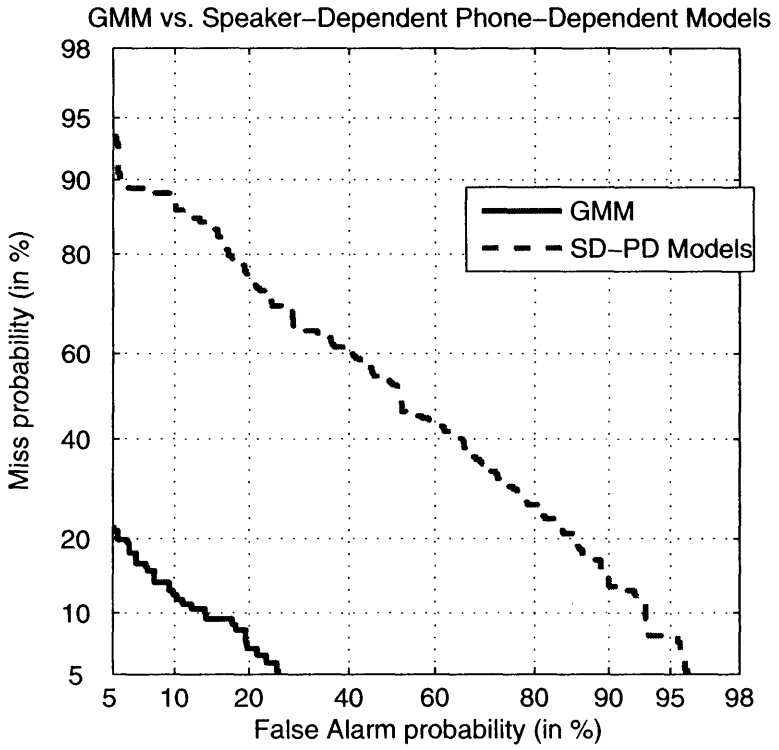

Figure 4-1: Detection error tradeoff curves for speaker-specific GMM and phone-dependent speaker-dependent models

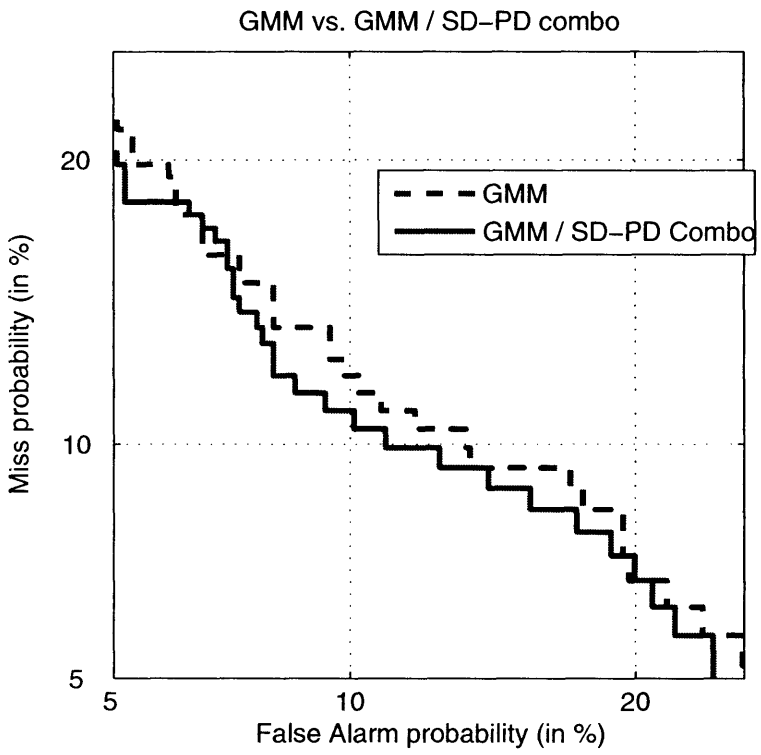

Figure 4-2: Detection error tradeoff curves for speaker-specific GMM and GMM/PD-SD combination models

4.1.3 Comparison of Landmark, Segment, & Frame Based Measurements

In modeling the speech signal, an acoustic-phonetic representation of the speaker can be based upon either a landmark, frame, or segment based framework. While landmark-based systems (as mentioned in 4.1.1) focus on acoustic boundaries, the segment-based framework extracts feature vectors from hypothesized variable-length phonetic segments, defined as the region between two landmarks. These feature vectors contain energy, duration as well as average and derivative Mel-frequency cepstral coefficient (MFCC) information. On the other hand, our frame-based system computes feature vectors at regular 5ms time intervals and concatenated average MFCCs from 4 regions surrounding the frame. In order to examine which framework, or combination of frameworks, best models speaker specific information, a module was developed to combine scores from multiple score model types and classifiers used in parallel.

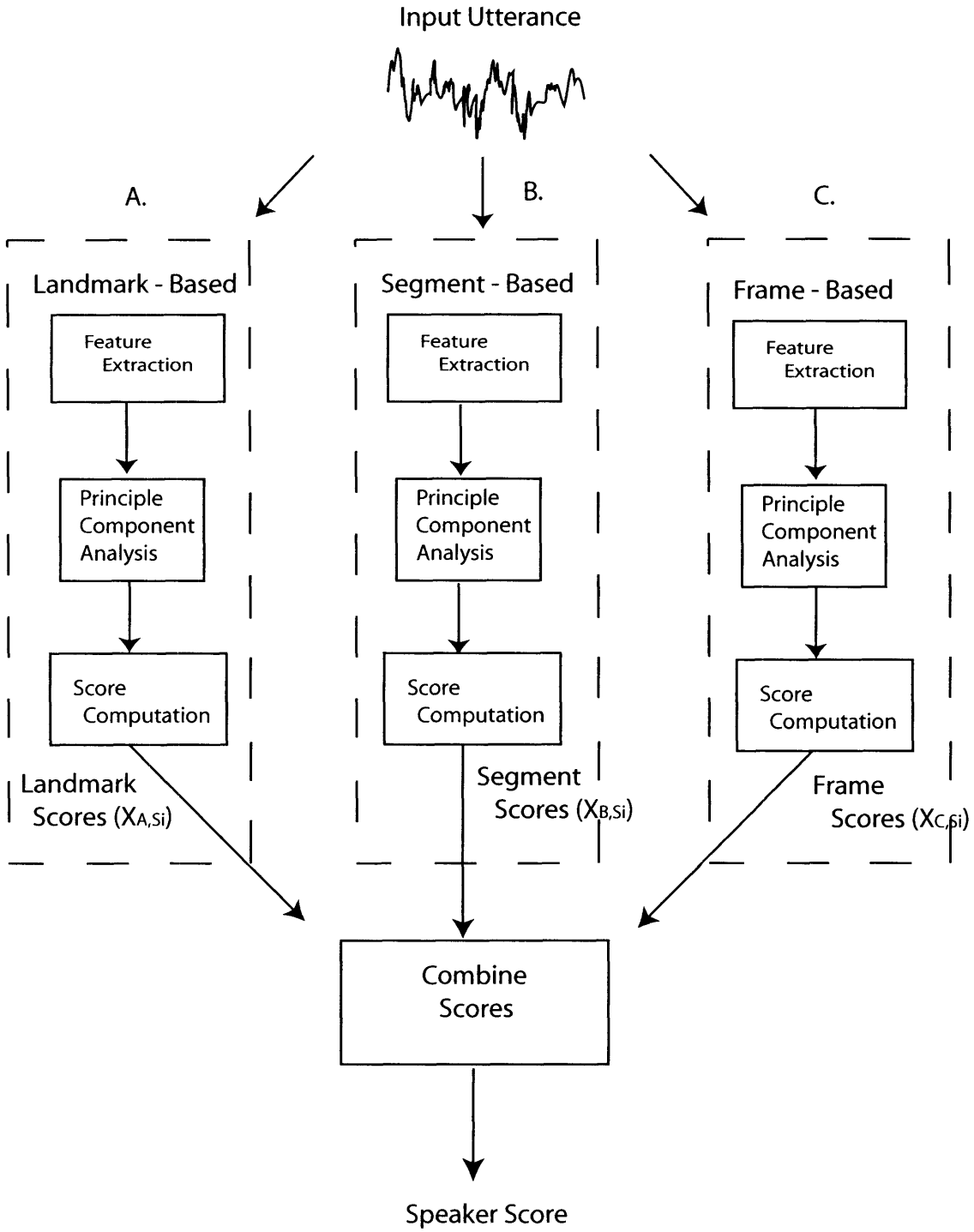

Figure 4-3: Block diagram of parallel landmark, frame, and segment based verification system

Figure 4-3 illustrates the process. The scores outputted from each independent classifier are combined to produce a combined speaker score, S_i:

$$S_i = \alpha x_{A,S_i} + \beta x_{B,S_i} + \gamma x_{C,S_i} \tag{4.3}$$

$$s.t. \quad \alpha + \beta + \gamma = 1 \tag{4.4}$$

In the following experiments, frame based models, segment based models, and landmark based models were trained for each enrollment speaker with training and testing conditions identical to the previous section. Table 4.2 shows EERs as the weights α, β, and γ are varied.

Landmarks: α	**Segment: β**	**Frames: γ**	**EER**
1	0	0	10.42%
0	1	0	11.46%
0	0	1	27.08%
0.3	0.7	0	9.99%
0.4	0.6	0	9.38%
0.5	0.5	0	10.24%
0.6	0.4	0	9.86%
0.7	0.3	0	10.42%

Table 4.2: EERs as landmark, segment, and frame-based scores are linearly combined in various ratios

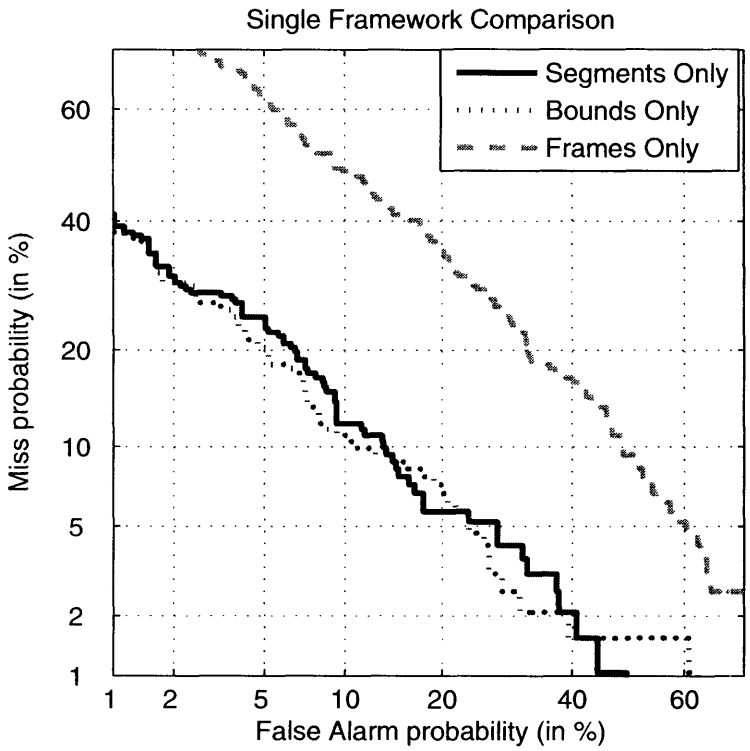

Figure 4-4: Detection error tradeoff curve for landmark only, segment only, and frame only models

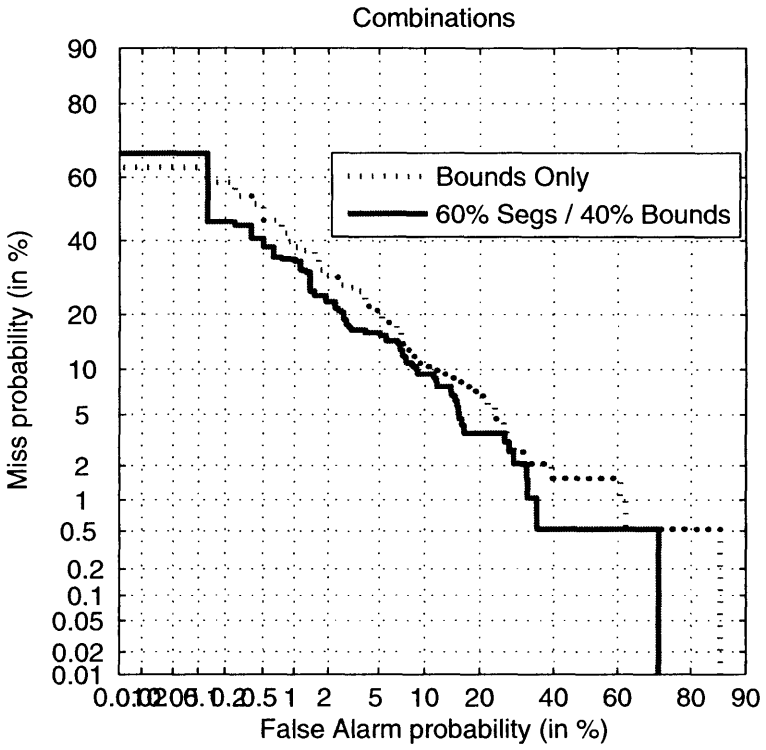

Figure 4-5: Detection error tradeoff curve for landmark only and 60% segment / 40% landmark weighted framework

When only results from a single model type were viewed, landmark based models produced the greatest verification accuracy. While models trained from segments produced similar, albeit slightly worse results, the verification accuracy of frame based models was particularly poor. This lackluster performance of frame based models sharply differs from previous experiments conducted [5]. We did not understand why frame based models produced such mediocre results and further investigation is needed.

Although boundary-only based models produced an EER of 10.42%, further improvements in performance were gleamed when the scores of all three model types (w/ $\alpha = 0.4$, $\beta = 0.6$, and $\gamma = 0$) were combined as seen in Figure 4-7. By combining the outputs from multiple classifiers, errors attributed to any one classifier were reduced in the final score, leading to increased verification accuracy. While moving from $\alpha = 0.4$, $\beta = 0.6$, and $\gamma = 0$ or $\alpha = 0.6$, $\beta = 0.4$ to $\alpha = 0.5$, $\beta = 0.5$, and $\gamma = 0$ produced what appeared to be a degradation in the EER (from 9.38% to 10.24%), Figure 4-6 reveals these differences to be mainly anomalous as the DET curves are nearly identical.

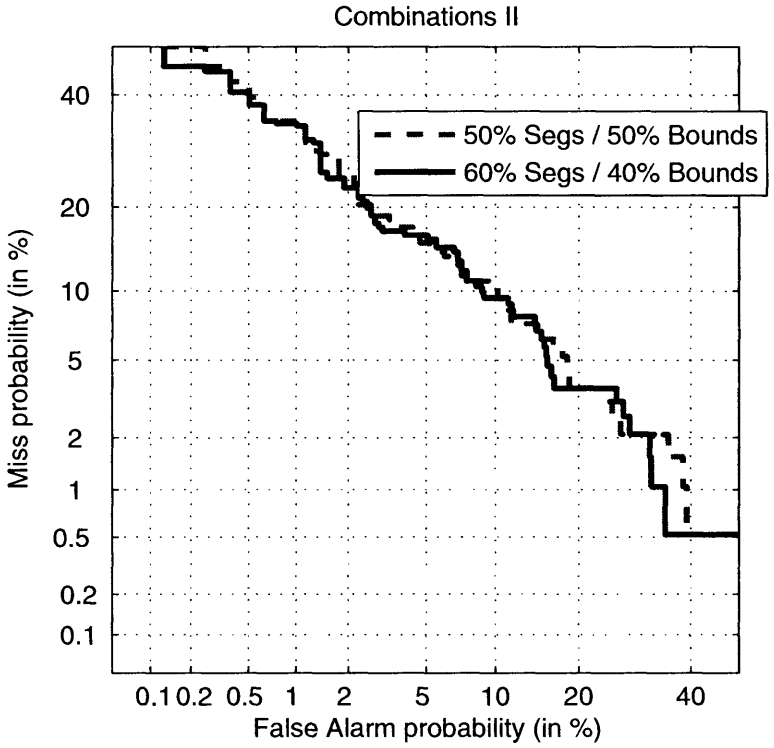

Figure 4-6: DET curves for 50% segment /50% landmark and 60% segment /40% landmark weighted frameworks

4.1.4 Mel-Frequency Cepstral Coefficients

Mel-frequency cepstral coefficients are one of the most dominant features used in speech recognition to model the spectral energy distribution of speech. When all coefficients are used, the Mel-scale speech spectrum is modeled exactly. However, as the number of MFCCs is reduced, the spectrum is gradually "smoothed", providing a model of the coarse spectral shape. Generally, speech recognizers only utilize the first 14 MFCCs as speech recognition is primarily concerned with the identification of formant locations. However, by only capturing the first 14 MFCCs, many speaker-specific characteristics important in speaker verification, such as formant bandwidth and fundamental frequency, are "smoothed" away. In order to understand the effects of Mel-frequency cepstral coefficients on speaker verification performance, we analyzed system performance as the number of MFCCs was varied from 10 to 26.

In general, as the number of MFCCs was increased from 14 to 24, system performance improved and the equal error rate (EER) decreased from 10.42% to a low of 8.85%. However, as the number of MFCCs was increased beyond 24, system performance began degrading. With the larger number of MFCCs leading to a less smoothed spectrum, it is believed that noise is a large contributor to the experienced performance decrease. Although 20 MFCCs produced the best EER, the resulting DET curve in Figure displayed undesirable characteristics in the lower right and upper left regions. Hence, we chose 24 MFCCs to be optimal.

MFCCs	EER
14	10.42%
16	10.42%
18	9.38%
20	8.85%
22	9.38%
24	9.38%
26	10.94%

Table 4.3: EERs as the number of Mel-frequency cepstral coefficients is varied from 10 to 26

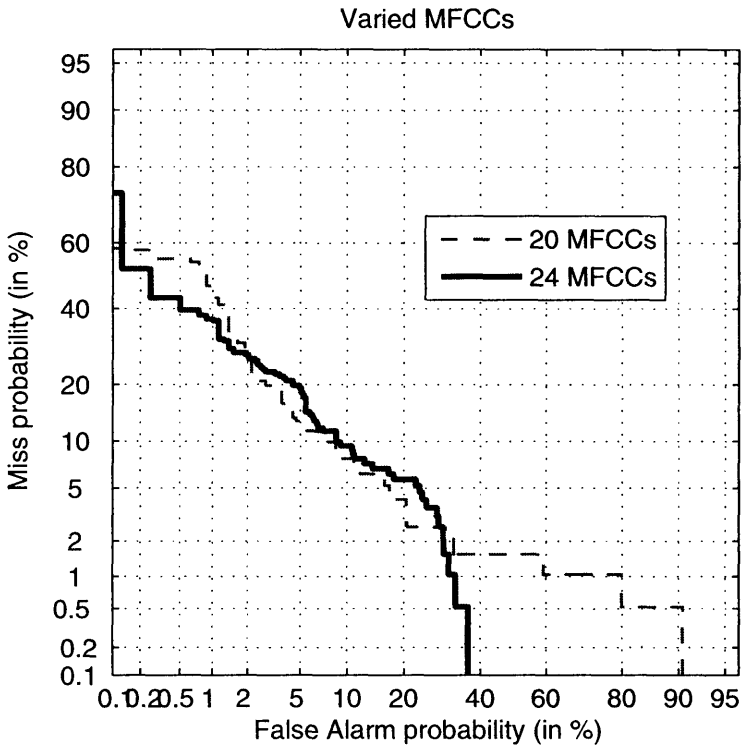

Figure 4-7: DET curves with the number of MFCCs equal to 20 and 24. Based upon a 60% segment / 40% landmark framework with $\tau = 5$

4.2 Experimental Conditions

In all of the following experiments, the speaker verification system extracted 24-dimension mean normalized MFCC feature vectors from each speech waveform, utilizing a 60% segment / 40% landmark based framework. Furthermore, the underlying system was an ASR-dependent speaker verification system coupled with phone adaptive normalization. The tuning factor, τ was set to 5, providing of combination of GMM / SD-PD model scores.

4.3 Effects of Mismatched Testing Conditions

In this section, experiments were conducted exploring the effects of mismatched testing conditions on system performance. In particular, we examined the impact of environment and microphone variability inherent with handheld devices. Figure 4-8 provides a preliminary glimpse of the impact of environment and microphone conditions. For these trials, known users enrolled by repeating a single ice cream phrase four times in a particular environment/microphone condition. During testing, both the enrolled user and dedicated imposter repeated the same ice cream flavor phrase. As can be seen, system performance varies widely as the environment or microphone is changed between the training and testing phase. While the fully matched trial (trained and tested in the office with an external earpiece headset) produced an EER of 9.38%, moving to a matched microphone/mismatched environment (trained in hallway/external, tested in intersection/external)resulted in a relative degradation of over 300% (EER of 29.17%). The following provide a greater in-depth analysis of these environment and microphone effects.

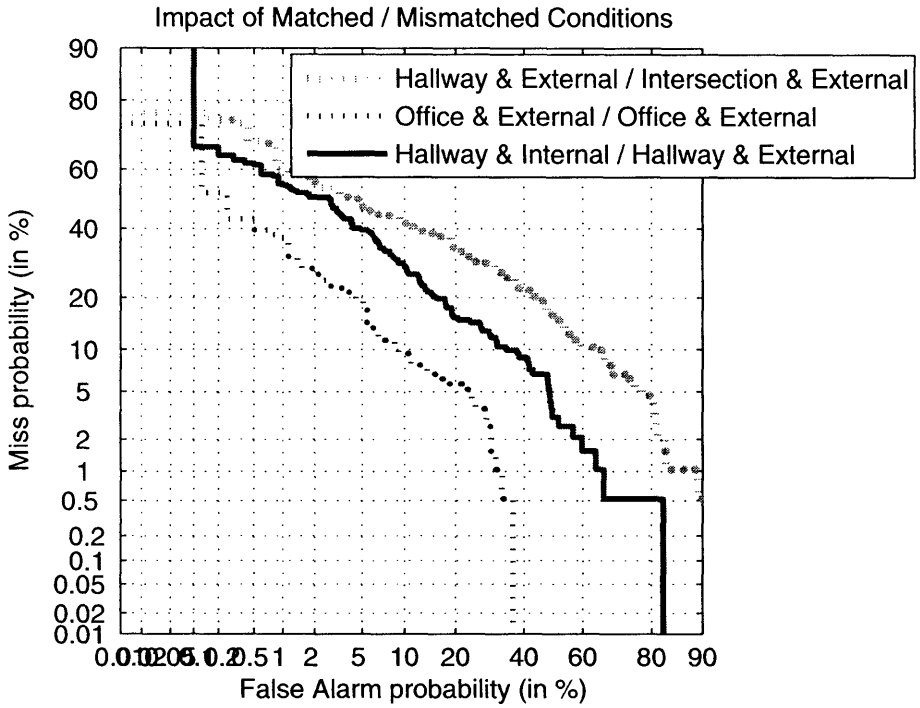

Figure 4-8: DET curves of preliminary cross-conditional tests with both matched and mismatched environment and microphone conditions.

4.3.1 Varied Environmental Conditions

As discussed in Section 1.2, the mobile nature of handheld devices exposes a speaker verification system to highly variable acoustic environments as well as background noises. In order to understand the effects of environment on speaker verification performance, we conducted a number of experiments. In each of the three trials, the speaker verification system was trained upon enrollment data collected in each of the following environments:

1. Office

2. Hallway

3. Intersection

	Trained on Office	Trained on Hallway	Trained on Intersection
Tested w/ Office	13.75%	13.33%	18.33%
Tested w/ Hallway	14.58%	14.79%	15.62%
Tested w/ Intersection	28.33%	30.00%	12.71%

Table 4.4: EERs of cross-conditional environment tests with models trained and tested in each of the three different environments leading to 9 distinct tests

Users enrolled by uttering five different name phrases two times each (once with both the headset and internal microphones) during the initial enrollment session[1]. System performance was then evaluated by testing the speaker verification system against data collected in each of the three environments. In all tests, the phrases used in the enrollment session were identical to the phrases in the testing session. This was fundamentally harder in comparison to the tests conducted in Section 4.1.4 as each name phrase is spoken only once for a given microphone/environment condition rather than 4 times. This is reflected in the higher EER of 13.75% seen in the train in office / test in office trial as opposed to the EER of 9.38% experienced when we trained and tested solely on a single phrase uttered in the office/external condition. These results from our tests are compiled in Table 4.4:

Several interesting observations can be made from these results. In general, one would expect that the speaker verification system would have the lowest equal error rates (EER) in situations where the system is trained and tested in the same environmental conditions. However, when the speaker verification system was trained in the hallway environment, the system performed better when tested in the office (13.33%) as opposed to the hallway environment (14.79%). Next, when trained in the intersection environment, the speaker verification system proved most robust with a maximum performance degradation of 5.65% as compared to 14.58% and 16.67% for office and hallway trained models. Furthermore, the train-intersection / test-

[1]Names, rather than ice cream flavor phrases, were used as examples as each name phrase appeared in all of the six conditions while ice cream flavors each appeared in only one condition for a given phrase list. This limited the number of matched/mismatched environment and microphone tests that could be achieved with ice cream flavor phrases.

intersection trial produced the lowest overall EER of 12.71%. This high performance factor could possibly be attributed to the varied background noise experienced in the intersection environment leading to speaker models that are more robust to noise. Overall, it appears that the performance degradation experienced when moving from a "noisy" training environment to a "clean" testing environment was not as drastic as that of the reverse situation.

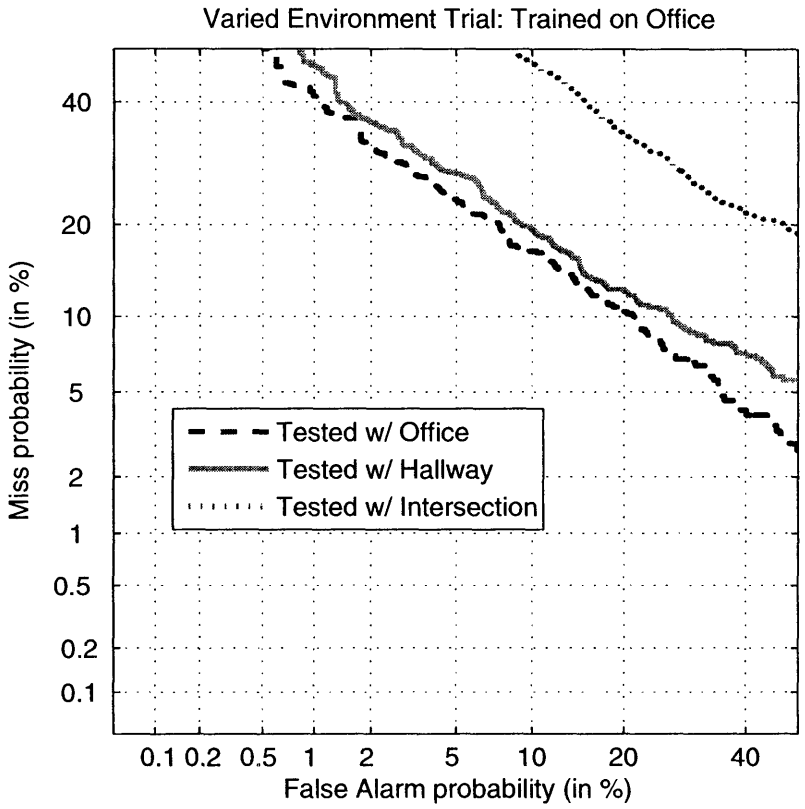

Figure 4-9: DET curve of models trained on name phrases in the office environment and tested in the three different environments (office, hallway, intersection)

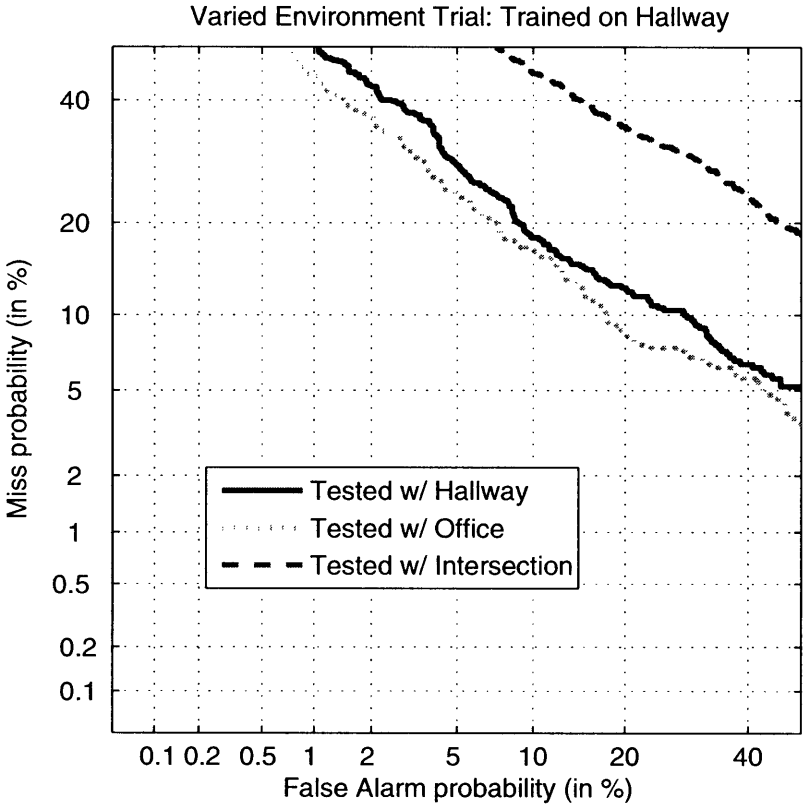

Figure 4-10: DET curve of models trained on name phrases in the hallway environment and tested in the three different environments (office, hallway, intersection)

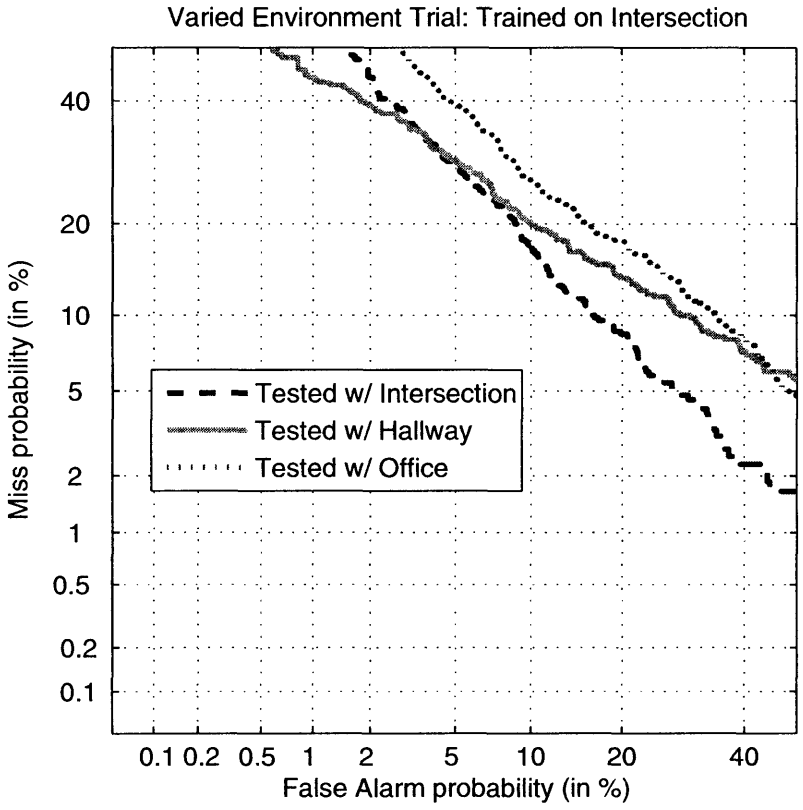

Figure 4-11: DET curve of models trained on name phrases in the hallway environment and tested in the three different environments (office, hallway, intersection)

4.3.2 Varied Microphone Conditions

Along with varied environmental conditions, speaker verification systems for handheld mobile devices are subjected to varying microphone conditions as a number of headset microphones can be used interchangeably with these devices. In order to understand the effect of microphones on speaker verification performance, we conducted a number of experiments in which the system was trained from data collected with either the internal microphone or an external headset. Therefore, users enrolled by uttering five different name phrases three times each (once in each of the environment conditions) during the initial enrollment session. Subsequently, the trained system was then tested on data collected in both conditions. The experimental conditions were identical to that of Section 4.2. The results of these trials can be seen in Table 4.5. From these results, it can be seen that varying the microphone used can have a huge impact on system performance. In both cases, if the system was trained and tested using the same microphone, the EER was approximately 11%. However, if the system was trained and tested using different microphones, we see a performance degradation of almost 8% - 11%. In terms of overall performance, it appears that training with the internal microphone leads to the best results.

	Trained on External	Trained on Internal
Tested w/ External	11.11%	18.19%
Tested w/ Internal	22.36%	10.97%

Table 4.5: EERs of cross-conditional microphone tests with models trained and tested with each of the two microphones (external and internal) leading to 4 distinct tests

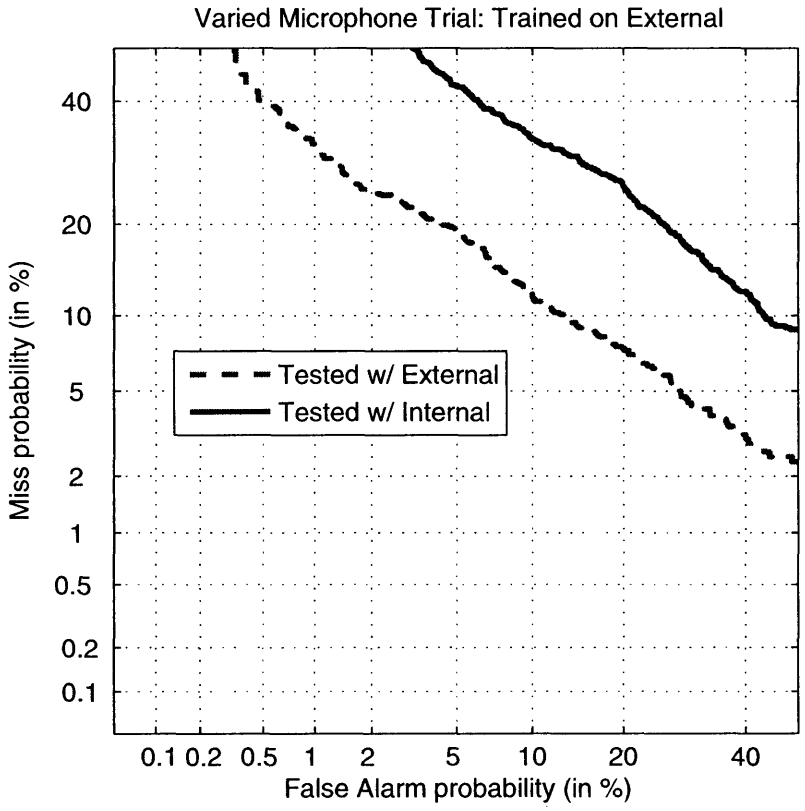

Figure 4-12: DET curve of models trained on name phrases with the handset microphone and tested with two different microphones (external and internal)

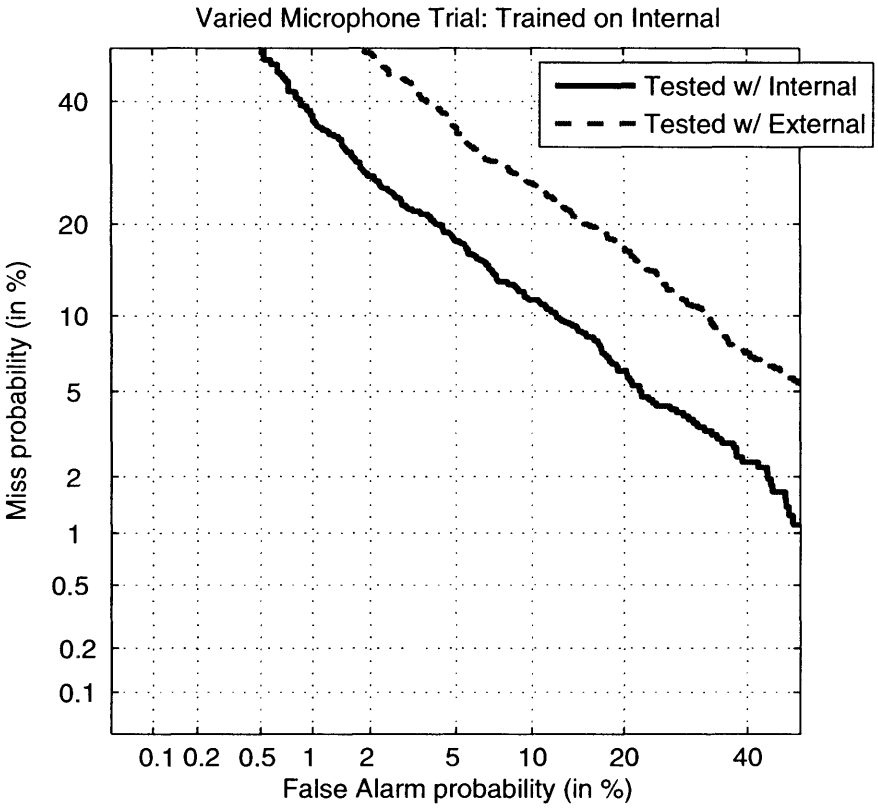

Figure 4-13: DET curve of models trained on name phrases with the internal microphone and tested with two different microphones (external and internal)

4.4 Methods for Improving Robustness

As previously illustrated, environment and microphone variabilities introduce severe challenges to speaker verification accuracy. This section describes three methods, handset dependent score normalization, zero normalization, and multistyle training, used to minimize degradations introduced by these factors.

4.4.1 Handset Dependent Score Normalization (H-norm)

Microphones introduce channel effects which create distortions in a user's speech signal. Hence, speaker models reflect not only speaker-specific characteristics, but also capture the characteristics of the microphone used [14]. The handset normalization technique (H-norm), developed by Reynolds, seeks to decouple the effects of the channel from the speech signal using speaker-specific handset statistics:

$$S_{HNORM}(X|s) = \frac{S(X|s) - \mu_s(mic)}{\sigma_s(mic)} \tag{4.5}$$

where $\mu_s(mic)$ and $\sigma_s(mic)$ are respectively the mean and standard deviation of a speaker model's scores to development set speech utterances captured with that particular microphone. Note that the development set does not contain speech from the enrolled users nor the dedicated imposters.

For our experiments, the development set was created by removing half of the speakers from the dedicated imposter set. Experiments were identical to Section 4.3.2 with the only difference being half of the imposters was removed for use in a development set. Tables 4.6 and 4.7 show a comparison of matched / mismatched

	Trained on External	Trained on Internal
Tested w/ External	10.42%	18.33%
Tested w/ Internal	21.11%	10.42%

Table 4.6: Unnormalized EERs from cross-conditional microphone tests with models trained and tested with two different microphones.

	Trained on External	Trained on Internal
Tested w/ External	9.44%	17.22%
Tested w/ Internal	14.86%	10.00%

Table 4.7: EERs after handset normalization from cross-conditional microphone tests, with models trained and tested with two different microphones

microphone tests with and without the use of H-norm.

There are two major trends which to note. First, H-norm reduced EER in all situations. The greatest improvement in accuracy occurred in the mismatched microphone trials with an absolute reduction of 6.25% for the trained w/ external tested on internal condition and 1.11% reduction for the trained w/ internal tested on headset condition. The second major trend is that all normalized DET curves appear to be a clockwise rotated version of their unnormalized counterparts as can be seen from Figures 4-14 to 4-17.

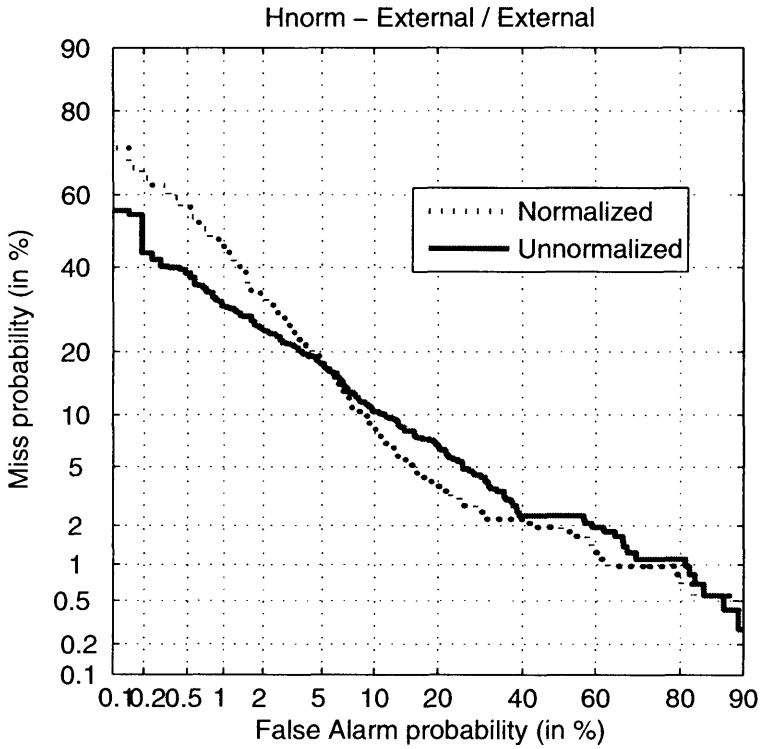

Figure 4-14: Unnormalized and normalized (H-norm) DET curves with models trained with the headset microphone and tested with the headset microphone

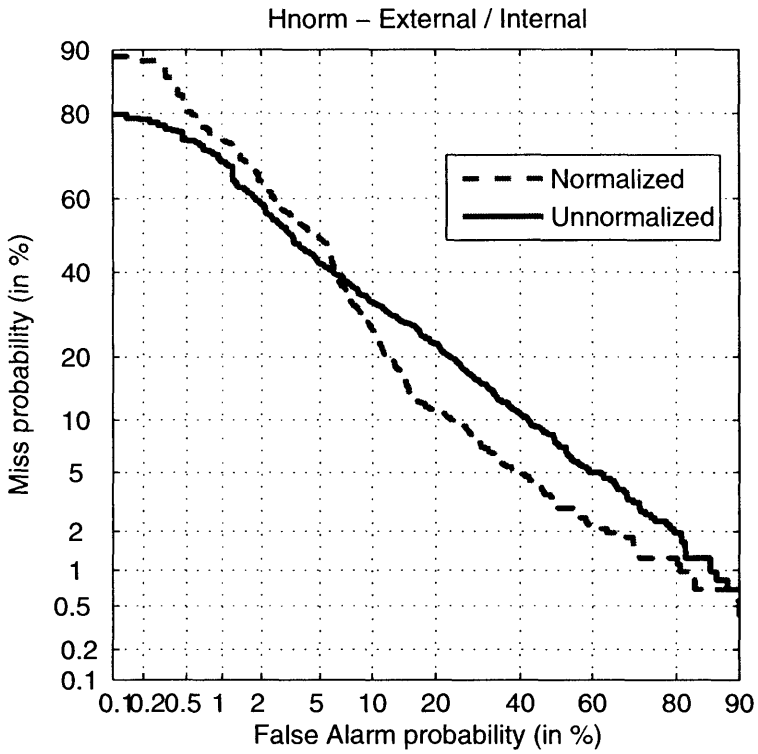

Figure 4-15: Unnormalized and normalized (H-norm) DET curves with models trained with the headset microphone and tested with the internal microphone

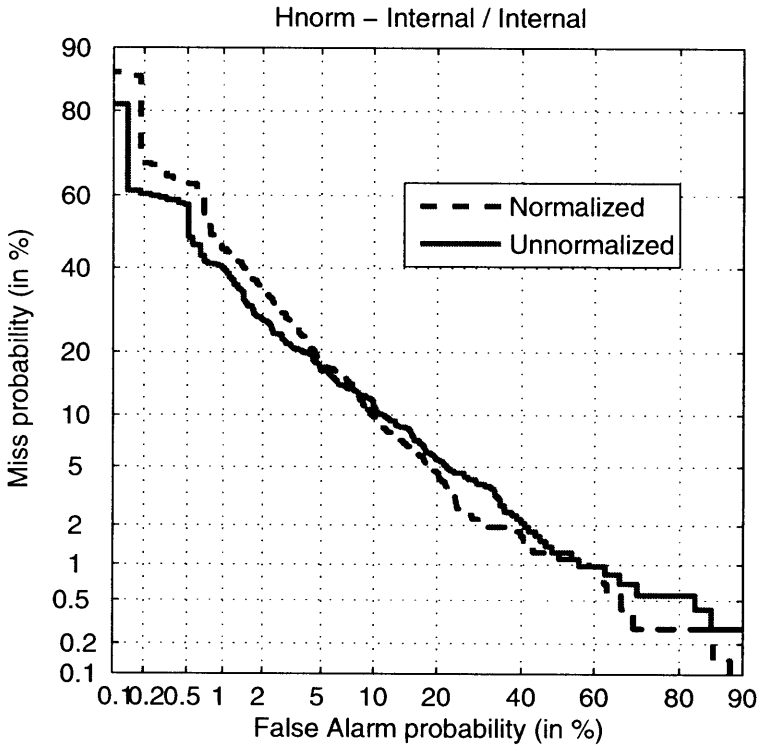

Figure 4-16: Unnormalized and normalized (H-norm) DET curves with models trained with the internal microphone and tested with the internal microphone

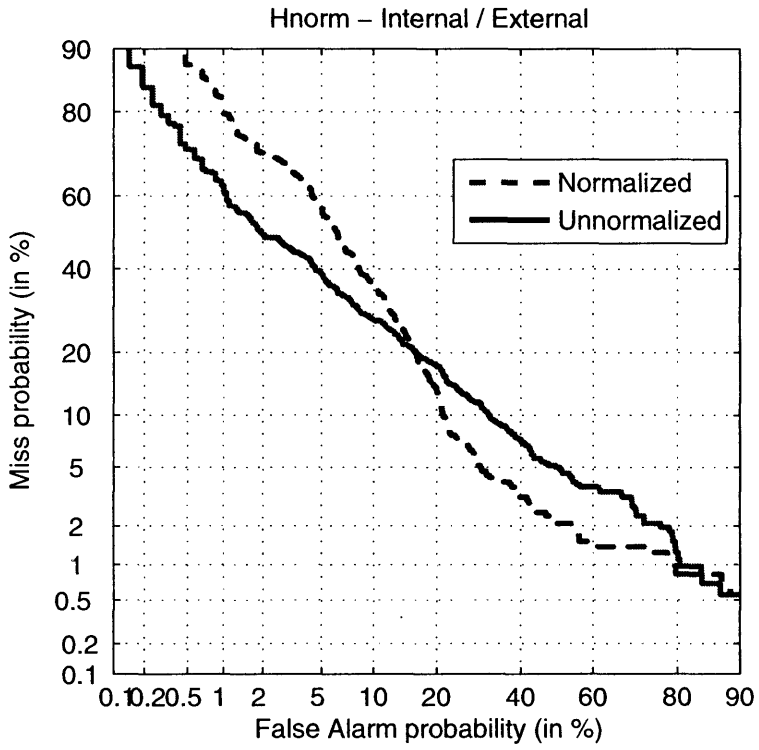

Figure 4-17: Unnormalized and normalized (H-norm) DET curves with models trained with the internal microphone and tested with the headset microphone

4.4.2 Zero Normalization (Z-norm)

As seen in Section 4.5, the H-norm technique can produce significant reductions in errors for mismatched microphone conditions. However, this technique is heavily reliant on the ability to accurately determine microphone labels for both development and test set utterances. Not only are the microphone labels necessary to create microphone-specific statistics, they also affect whether speaker scores are correctly normalized by the appropriate statistics.

A closely related technique, Zero Normalization (Z-norm), provides both microphone and speaker normalization while bypassing these aforementioned difficulties. Z-norm can be described as [3]:

$$S_{znorm}(X|s) = \frac{S(X|s) - \mu_s}{\sigma_s} \tag{4.6}$$

where μ_s and σ_s are respectively the mean and standard deviation of a speaker model's scores to all development set speech utterances regardless of the microphone. Hence, the Z-norm procedure proved simpler than the H-norm technique.

Once again, the development set was created by removing half of the speakers from the dedicated imposter set. Experimental conditions were identical to Section 4.5. Tables 4.8 and 4.9 show a comparison of matched / mismatched microphone tests with and without the use of Z-norm.

	Trained on External	Trained on Internal
Tested w/ External	10.42%	18.33%
Tested w/ Internal	21.11%	10.42%

Table 4.8: Unnormalized EERs from cross-conditional microphone tests with models trained and tested with two different microphones.

	Trained on External	Trained on Internal
Tested w/ External	9.44%	15.42%
Tested w/ Internal	15.32%	11.25%

Table 4.9: EERs after zero normalization (Z-norm) from cross-conditional microphone tests, with models trained and tested with two different microphones

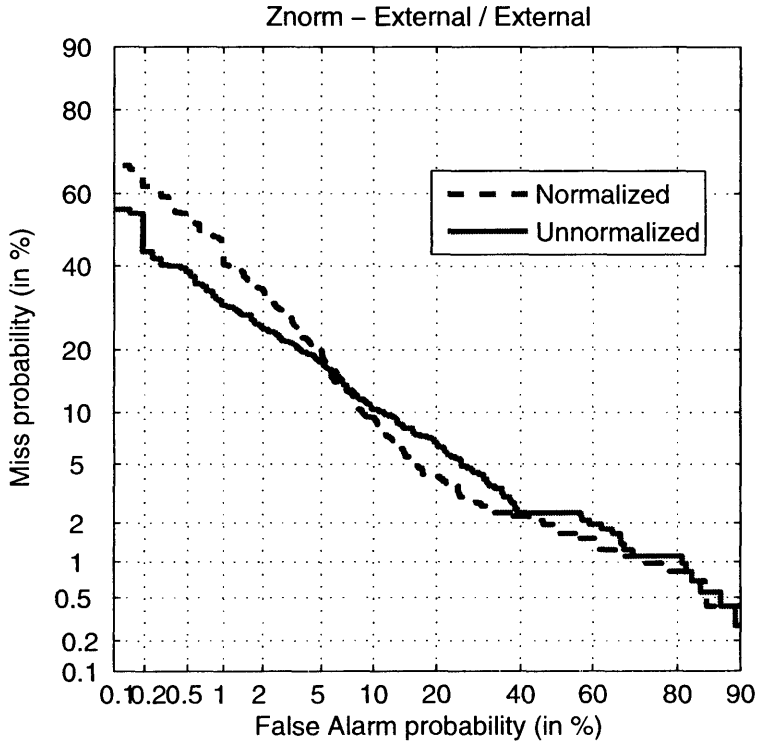

Figure 4-18: Unnormalized and normalized (Z-norm) DET curves with models trained with the headset microphone and tested with the headset microphone

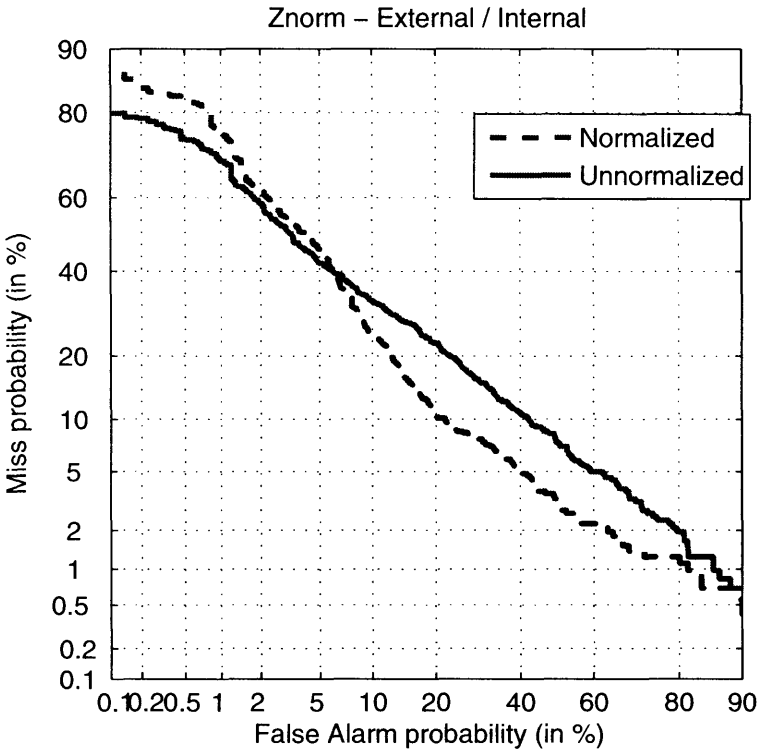

Figure 4-19: Unnormalized and normalized (Z-norm) DET curves with models trained with the headset microphone and tested with the internal microphone

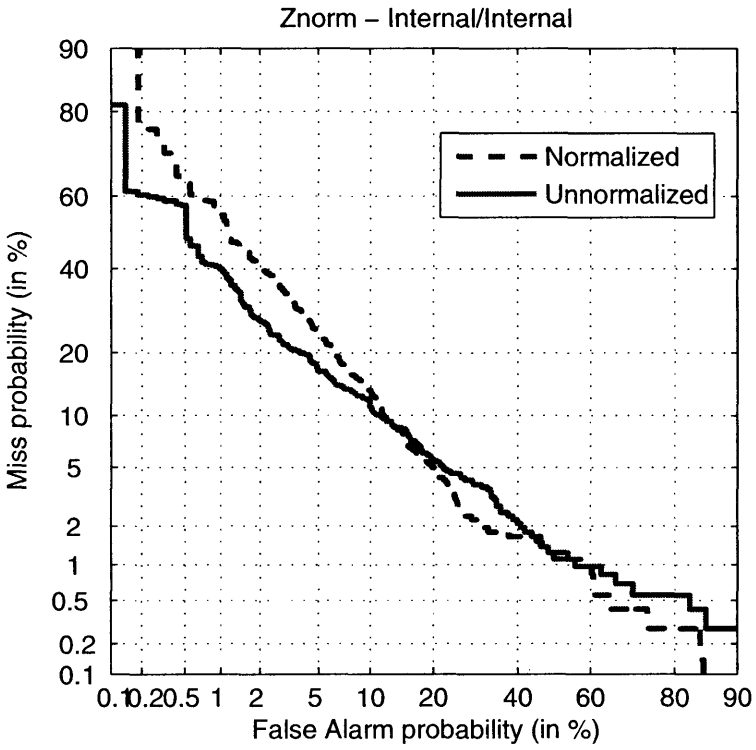

Figure 4-20: Unnormalized and normalized (Z-norm) DET curves with models trained with the internal microphone and tested with the internal microphone

As can be seen, the Z-norm technique can produce significant reductions in errors for the mismatched microphone conditions. Although, in general, these improvements in performance lag that seen with H-norm, Z-norm requires less information about each speech utterance.

4.4.3 Multistyle Training

While H-norm and Z-norm attempt to improve speaker verification accuracy by decoupling the effects of the microphone from the speech signal through post-processing (after the models have been created), multistyle training takes a different track and works to improve the underlying speaker models. For multistyle training, the enrolled user recorded a single name phrase in each of the 6 testing conditions, essentially sampling all possible environment and microphone conditions. Therefore, rather than training highly focused models for a particular microphone or environment, multistyle training develops diffuse models which cover a range of conditions. These models were then tested against imposter utterances from particular microphone or environment conditions with the results shown below:

Tested in office	7.77%
Tested in hallway	10.01%
Tested in intersection	12.92%
Tested in all locs/mics	11.11%

Table 4.10: EERs of multistyle trained models tested in three different locations

Tested with external	8.13%
Tested with internal	9.67%
Tested in all locs/mics	11.11%

Table 4.11: EERs of multistyle trained models tested with two different microphones

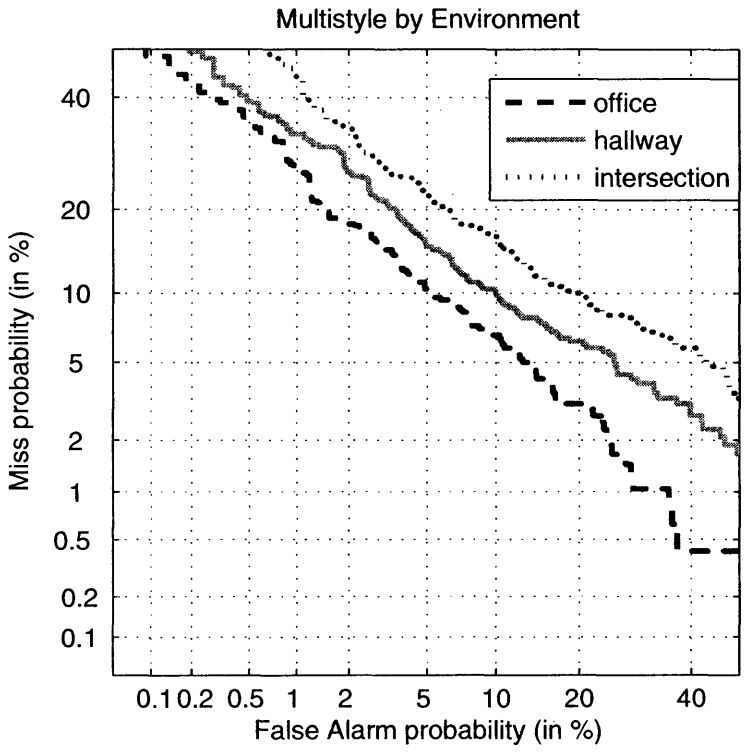

Figure 4-21: DET curves of multistyle trained models tested in three different locations

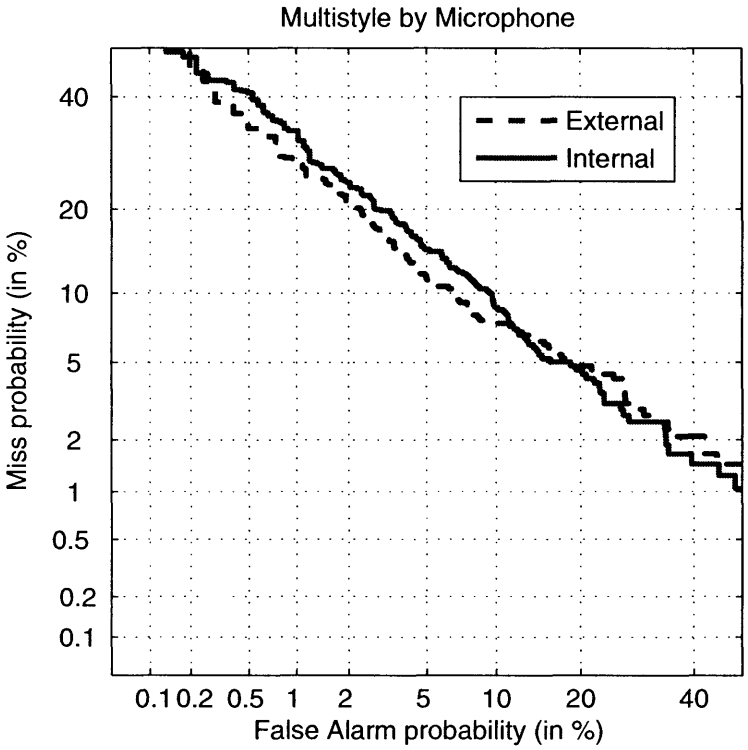

Figure 4-22: DET curves of multistyle trained models tested with two different microphones

Despite only being trained on 6 enrollment utterances, multistyle models performed better than models trained solely in one environment or with a single microphone but with a greater number of speech utterances (10 to 15) as seen by comparing Tables 4.4 and 4.5 to Tables 4.10 and 4.11. Furthermore, multistyle models appear more resilient to performance degradations caused by changing microphones or environments. When comparing maximum performance degradations, multistyle models experienced an absolute decrease in accuracy of 5.149% when moving from testing in the best environment to the worst (i.e. in this case from the office to the intersection). Cross-conditional tests, however, experienced maximum performance degradations of 14.58%, 16.67%, and 5.62% when trained in the office, hallway, and intersection environments, respectively. Likewise, similar results hold when comparing across microphone conditions. This indicates that having at least a small amount of data from each environment / microphone can significantly improve performance and robustness.

4.5 Knowledge

In this section, we explore how knowledge of the correct log-in passphrase affects a speaker verification system's ability to correctly discriminate the "true" user from imposters.

4.5.1 Impact of Imposter's Knowledge of Passphrase

Although speaker verification seeks to provide security through a user's voice characteristics, we explored whether the application of random user selected login passphrases could provide an additional layer of security. Under this scenario, rather than prompting users to read openly displayed phrases, system users are asked to recite a secret user-specific passphrase chosen during the enrollment session. In our research, we conducted multistyle tests, under the same experimental conditions as Section 4.4.3 which did not explicitly verify the accuracy of the spoken passphrase, focusing only on speaker voice characteristics. However, in one test all enrolled users attempted to log-in with the correct passphrase while dedicated imposters spoke a variety of mostly incorrect phrases. This mimics the situation where an unknowledgeable imposter attempts to gain system access by randomly guessing passphrases, occasionally hitting upon the correct one. During the speech recognition component, incorrect spoken utterances (i.e. not the correct passphrase) were correctly aligned rather than forcibly aligned to what the correct passphrase should be. In a second test we conducted, both the enrolled users and imposters attempted to log-in with full knowledge of the correct passphrase. Figure 4-24 shows the results of these experiments.

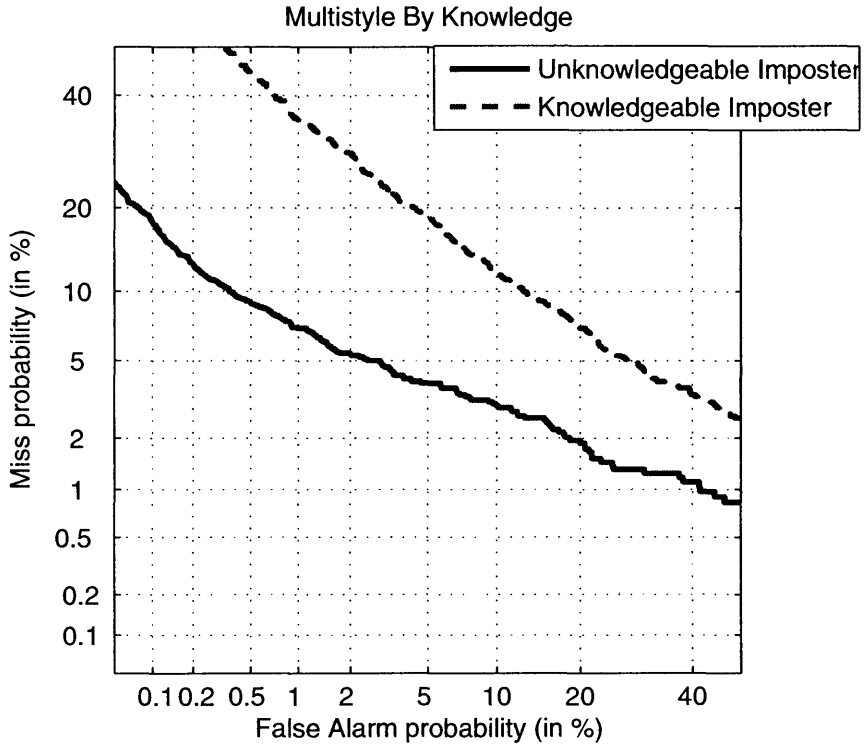

Figure 4-23: DET curves for multi-style trained models tested under the condition that the imposters either have or do not have knowledge of the user's passpharse.

As can be seen, the EER dramatically improves from 11.11% to 4.1% when imposters do not have knowledge of the user's passphrase. Hence, the use of secret passphrases can provide enormous benefit in discriminating enrolled users from imposters. This improvement is attributed to the speaker-specific GMM as SD-PD models trained from a single passphrase would likely contain few, if any, phone-level models for phones found in an incorrect utterance. While the relative 63% reduction in EER is impressive, additional methods provided further improvement. One possible method we explored was to completely reject any speaker whose utterance did not match the correct passphrase rather than proceeding with verification on the incorrect utterance. This eliminated all but the most dedicated imposters and produced an EER of 1.25%. Furthermore, by rejecting all unknowledgeable imposters outright, the maximum false acceptance rate was greatly reduced to 2%.

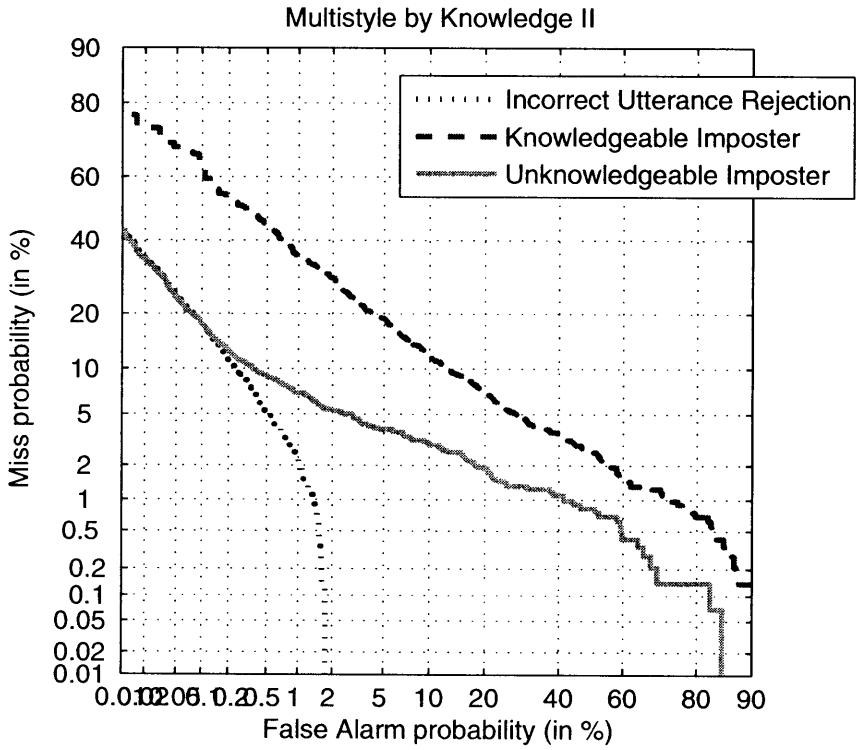

Figure 4-24: DET curves comparing multi-style trained models in which all unknowledgeable imposters are rejected outright

Another possible approach for future work in reaping further improvements would be to forcibly align incorrect utterances to the correct passphrase during speech recognition. This incorrect alignment should result in scores lower than for correctly aligned utterances.

Chapter 5

Conclusions

5.1 Summary

Throughout this thesis, we explored the problem of robust speaker verification for handheld devices under the context of extremely limited training data. This work analyzed basic speaker verification modeling techniques, the effects of mismatched testing conditions, methods for improving robustness, as well the impact of knowledge on verification accuracy.

5.1.1 Basic Speaker Verification Modeling

In Section 4.1, we explored a number of basic speaker verification modeling techniques. We first compared whether, speaker-dependent global GMMs or speaker-dependent phone-dependent models best captured speaker specific acoustic information. As sparse enrollment data prevented the training of robust models at the phone level, GMMs proved superior to SD-PD models in our experiments. However, additional improvements were made possible by combining phone dependent and GMM speaker model probabilities. This technique utilized SD-PD models only when a robustly trained phone model existed, otherwise backing off to the speaker GMM.

The second modeling technique explored centered around feature extraction. In

modeling the speech signal, an acoustic-phonetic representation of the speaker can be based upon either landmark, frame, or segment based features. When comparing the results from our experiments, we observed that while landmark-only based models provided the greatest accuracy of any single model type, combining scores from multiple model types proved most effective. By combining the outputs from multiple classifiers, errors attributed to any one classifier were reduced in the final score, leading to increased verification accuracy.

Finally, we analyzed the impact of Mel-frequency cepstral coefficients by conducting experiments varying the number of MFCCs from 10 to 26. As we increased the number of MFCCs, we found verification accuracy to initially improve, peaking around 22 MFCCs before slowly degrading. One possible reason for this is that, while utilizing fewer than approximately 20 MFCCs is insufficient to capture important speaker-specific characteristics, using greater than 24 MFCCs leads to a noisy Mel-scale speech spectrum.

5.1.2 Mismatched Testing Conditions

Section 4.3 discussed the impact of mismatched testing conditions on speaker verification accuracy. From these experiments, it was apparent that mismatches in microphone or environment conditions resulted in severe performance degradations. However, it appears that the performance degradation experienced when moving from a "noisy" training environment to a "clean" testing environment was not as drastic as that of the reverse situation. This is likely due to the fact that the varied background noise experienced in a "noisy" training environment led to speaker models that are more robust against noise.

5.1.3 Methods for Improving Robustness

In order to improve robustness against environment and microphone variabilities, Section 4.4 explored three methods, handset dependent score normalization, zero

normalization, and multistyle training to minimize degradations introduced by these factors. Both score normalization techniques, H-norm and Z-norm attempted to remove microphone-dependent and speaker-dependent biases from speaker scores. Our experimental results show significant reductions in EER, particularly in mismatched microphone conditions with the use of these techniques. Although H-norm provided greater improvements in performance, Z-norm benefited from a simpler implementation as no prior knowledge of microphone information was needed to develop speaker-specific statistics.

The third method we investigated, multistyle training, worked to improve the underlying speaker models by training diffuse models which sampled all possible environment and microphone conditions. This not only resulted in improved verification accuracy, multistyle trained models were also more resilient to performance degradations caused by changing microphones or environments.

5.1.4 Impact of Knowledge

Finally, in Section 4.5, we explored how knowledge of the correct log-in passphrase affects a speaker verification system's ability to correctly discriminate the "true" user from imposters. By allowing enrolled users to select random login passphrases which are kept secret as opposed to utilizing openly displayed phrases, the EER was cut in half. Further improvements were seen when we completely reject any speaker whose utterance did not match the correct passphrase rather than proceeding with verification on the incorrect utterance.

5.2 Future Work

In the future, there are a number of areas improvements we would like to pursue. Initially, we hope to investigate the cause of the lackluster performance of frame-based models seen in our experiments. We believe the observed results could be

greatly improved upon and would contribute to further improvements in EER when combined with segment and landmark based models.

Due to the promising results of H-norm and Z-norm, we would like to further explore score normalization techniques by analyzing a third common method known as T-norm.

Finally, based on the results of multistyle training, we would like to expand upon this to explore explicit noise compensation techniques such as parallel model combination [9] or universal compensation [10]. Furthermore, we also hope to investigate methods to synthesize multi-style models from single condition data.

Bibliography

[1] www.nuance.com.

[2] www.scansoft.com.

[3] C. Barras and J.L. Gauvain. Feature and Score Normalization For Speaker Verification of Cellular Data. *ICASSP*, 2003.

[4] F. Bimbot, M. Blomberg, L. Boves, G. Chollet, C. Jaboulet, B. Jacobs, J. Kharroubi, J. Koolwaaij, J. Lindberg, J. Mariethoz, C. Mokbel, and H. Mokbel. An Overview of the PICASSO Project Research Activities In Speaker Verification For Telephone Applications. *Eurospeech*, 1999.

[5] G. Doddington. Speaker recognition - identifying people by their voices. *Proc. of IEEE*, 73(11).

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

[7] T.J. Hazen, E. Weinstein, and A. Park. Towards robust person recognition on handheld devices using face and speaker identification technologies. *Proc. of Int. Conf. on Multimodal Interfaces*, pages 19–41, 2003.

[8] L.F. Lamel and J.L. Gauvain. Speaker verification over the telephone. *Speech Communication*, 31(2-3):141–154, 2000.

[9] M.Gales and S.Young. Robust continuous speech recognition using parallel model combination. *Trans. on Speech and Audio Processing*, 4:352–359, 1996.

[10] J. Ming, D. Stewart, and S. Vaseghi. Speaker identification in unknown noisy conditions - a universal compensation approach. *ICASSP*, 2005.

[11] P. R. Morin and J. Junqua. A Voice-Centric Multimodal User Authentication System for Fast and Convenient Physical Access Control. *Proc. of Multimodal User Authentication*, 2003.

[12] A. Park and T.J. Hazen. ASR Dependent Techniques for Speaker Identification. *Proc. ICSLP*, pages 2521–2524, 2002.

[13] A. Park and T.J. Hazen. A Comparison of Normalization and Training Approaches for ASR-Dependent Speaker Identification. *Interspeech*, 2004.

[14] D.A. Reynolds. Comparison of Background Normalization Methods for Text-Independent Speaker Verification. *Proc. Eurospeech-97*, 1997.

[15] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1-3):19–41, 2000.