TELEPHONE-BASED TEXT-DEPENDENT SPEAKER VERIFICATION

by

Osman Büyük

B.S., Electrical and Electronics Engineering, Bilkent University, 2003

M.S., Electrical and Electronics Engineering, Sabanci University, 2005

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Electrical and Electronics Engineering

Boğaziçi University

2011

TELEPHONE-BASED TEXT-DEPENDENT SPEAKER VERIFICATION

APPROVED BY:

Prof. Levent M. Arslan …………………………
(Thesis Supervisor)

Assoc. Prof. Burak Acar …………………………

Assist. Prof. Hakan Erdoğan …………………………

Assoc. Prof. Engin Erzin …………………………

Assoc. Prof. Murat Saraçlar …………………………

DATE OF APPROVAL: 09.06.2011

ACKNOWLEDGEMENTS

I would like thank my supervisor Prof. Levent Arslan for his guidance and creative ideas throughout this thesis. Without his support, this thesis would not be possible. I am very grateful since I had a chance to work in Sestek Inc. It was a life-time experience for me.

I would like to thank Dr. Hakan Erdoğan for introducing me to speech recognition topic during my master study at Sabanci University. I feel very fortunate that I have had the opportunity to meet with him.

I would like to express my sincere gratitude to Dr. Murat Saraçlar. His valuable comments and friendly support always helped me to guide my research in the right direction.

I would like to thank Dr. Engin Erzin and Dr. Burak Acar for spending their valuable time and energy in my final presentation committee.

I owe special thanks to Prof. M. Ufuk Çağlayan for the financial support from the TAM project in the last two years of my study. This work is supported in part by the Turkish State Planning Organization (DPT) under the TAM Project, number 2007K120610.

I would like to thank my wife Melissa and my family Emel-Kadir Büyük, Olcay-Cihat İnceoglu, Sibel and Nadir for their endless support and patience. I would like to dedicate this thesis to my grandmother who passed away a few days before I finished writing this thesis.

Last but not least I would like thank to all my friends from Bogazici University, Sabanci University, Bilkent University and Sestek for their technical and non-technical contributions during this difficult process.

ABSTRACT

TELEPHONE-BASED TEXT-DEPENDENT SPEAKER VERIFICATION

In this thesis, we investigate model selection and channel variability issues on telephone-based text-dependent speaker verification applications. Due to the lack of an appropriate database for the task, we collected two multi-channel speaker recognition databases which are referred to as text-dependent variable text (TDVT-D) and text-dependent single utterance (TDSU-D). TDVT-D consists of digit strings and short utterances in Turkish and TDSU-D contains a single Turkish phrase.

In the TVDT-D, Gaussian mixture model (GMM) and hidden Markov model (HMM) based methods are compared using several authentication utterances, enrollment scenarios and enrollment-authentication channel conditions. In the experiments, we employ a rank-based decision making procedure. In the second set of experiments, we investigate three channel compensation techniques together with cepstral mean subtraction (CMS): i) LTAS filtering ii) MLLR transformation iii) handset-dependent rank-based decision making (H-rank). In all three methods, a prior knowledge of the employed channel type is required. We recognize the channels with channel GMMs trained for each condition. In this section, we also analyze the influence of channel detection errors on the verification performance.

In the TDSU-D, phonetic HMM, sentence HMM and GMM based methods are compared for the single utterance task. In order to compensate for channel mismatch conditions, we implement test normalization (T-norm), zero normalization (Z-norm) and combined (i.e., TZ-norm and ZT-norm) score normalization techniques. We also propose a novel combination procedure referred to as C-norm. Additionally, we benefit from the prior knowledge of handset-channel type in order to improve the verification performance. A cohort-based channel detection method is introduced in addition to the classical GMM-based method. After the score normalization section, feature domain spectral mean division (SMD) method is presented as an alternative to the well-known CMS. In the last set of experiments, prosodic (energy, pitch, duration) and spectral features are combined together in the sentence HMM framework.

ÖZET

TELEFON ÜZERİNDEN METNE BAĞIMLI KONUŞMACI ONAYLAMA

Bu tezde telefon üzerinden metne bağımlı bir konuşmacı tanıma uygulamasında model seçimi ve kanal değişkenliği konuları incelenmektedir. Çalışma için uygun Türkçe bir veritabanının bulunmaması nedeniyle, metne bağımlı değişken metin (MBDM) ve metne bağımlı tek cümle (MBTC) isimlerinde çok kanallı iki veritabanı toplanmıştır. MBDM veritabanı sayı dizileri, kısa cümle ve kelimelerden oluşurken, MBTC veritabanı tek bir cümleden oluşmaktadır.

MBDM veritabanında, Gauss karışım model (GKM) ve saklı Markov model (SMM) tabanlı iki metot farklı test cümleleri, eğitim senaryoları ve test-eğitim kanal durumları için karşılaştırılmıştır. Deneylerde sıralama tabanlı bir karar verme yöntemi kullanılmıştır. İkinci deney setinde, kanal uyumsuzluğu problemini gidermek için kepstral ortalama çıkarımı (CMS) ile birlikte, üç farklı yöntem denenmiştir: i) uzun dönemli ortalama spektrum (LTAS) filtrelemesi ii) maksimum olasılık doğrusal regresyon (MLLR) dönüşümü iii) kanal bağımlı sıralama tabanlı karar verme yöntemi (H-rank). Her üç yöntemde de, kullanılan kanal çeşidinin bilinmesi gerekmektedir. Kanal çeşidi her kanal için eğitilmiş kanal GKM'leri ile tanınmıştır. Deneylerde kanal tanıma hatalarının konuşmacı tanıma performansı üzerindeki etkisi de incelenmiştir.

MBTC veritabanında, cümle SMM, fonetik SMM ve GKM yöntemleri tek cümle uygulaması için karşılaştırılmıştır. Kanal etkisini gidermek için test normalizasyonu (T-norm), sıfır normalizasyonu (Z-norm) ve kombinasyonları (ZT-norm ve TZ-norm) gibi farklı skor normalizasyon metotları denenmiştir. Bu metotlara ek olarak C-norm isimli bir kombinasyon önerilmiştir. Başarımı arttırmak için skor normalizasyonu sırasında kanal bilgisinden de faydalanılmıştır. Kohort tabanlı bir kanal tanıma yöntemi klasik GKM yöntemine ek olarak denenmiştir. Skor normalizasyonu bölümünden sonra, spektral ortalama bölümü (SMD) yöntemi sık kullanılan CMS'ye bir alternatif olarak önerilmiştir. Bu veritabanındaki son deneylerde, spektral özellikler ile enerji, tonlama ve süre özellikleri cümle SMM yapısı içerisinde birleştirilmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$\boldsymbol{A}$	MLLR adaptation regression matrix
$\boldsymbol{b}$	MLLR adaptation bias vector
$d(\cdot,\cdot)$	Distance metric
$\boldsymbol{o}_t$	Observation vector at time *t*
$\boldsymbol{O}$	Sequence of observation vectors
q_t	HMM state at time *t*
$\boldsymbol{r}_t$	Reference vector at time *t*
w	Mixture weight
$\boldsymbol{\Sigma}$	Covariance matrix
$\boldsymbol{\mu}$	Mean vector
$\boldsymbol{\sigma}$	Standard deviation vector
μ	Mean value
σ	Standard deviation value
λ_{BKG}	Background model
λ_S	Speaker model
$\Lambda(\boldsymbol{O})$	Speaker score
$\Lambda_{norm}(\boldsymbol{O})$	Normalized speaker score

LIST OF ACRONYMS/ABBREVIATIONS

ANN	Artificial Neural Networks
BW	Baum-Welch
C-norm	Combined Normalization
CMN	Cepstral Mean Normalization
CMS	Cepstral Mean Subtraction
DCT	Discrete Cosine Transform
DET	Detection Error Tradeoff
DTW	Dynamic Time Warping
EER	Equal Error Rate
EM	Expectation Maximization
FA	Factor Analysis
FM	Feature Mapping
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
LDC	Linguistic Data Consortium
LPCC	Linear Prediction Cepstral Coefficients
LSF	Line Spectral Frequencies
LTAS	Long-Term Average Spectrum
MAP	Maximum A-Posterior
MFCC	Mel-Frequency Cepstral Coefficients
MLLR	Maximum Likelihood Linear Regression
NIST	National Institute of Standards and Technology
OLA	Overlap and Add
PLP	Perceptual Linear Predictive

ROC	Receiver Operating Characteristic
SI	Speaker Independent
SMD	Spectral Mean Division
SMS	Speaker Model Synthesis
STFT	Short Time Fourier Transform
SVM	Support Vector Machine
TDSU	Text-Dependent Single Utterance
TDVT	Text-Dependent Variable Text
T-norm	Test Normalization
UBM	Universal Background Model
VQ	Vector Quantization
Z-norm	Zero Normalization

1. INTRODUCTION

1.1. Motivation

Biometrics is the science of recognizing a person using his/her intrinsic physiological and/or behavioral traits. Due to increasing demand for secure applications and advent in new technologies, biometric market has been one of the fastest growing markets for the past few years (RNCOS, 2011). According to (BCC, 2010), compound annual growth rate of the market in the period between 2010 and 2015 is expected to be 18.9%.

The most common biometric techniques include the automatic recognition of fingerprint, face, iris, retina, hand geometry, signature and voice. Biometric identification systems (BISs) are employed in real-life applications depending on their characteristics such as reliability, ease of use, ease of implementation, user acceptance and cost in addition to the requirement of the final product. Table 1.1 provides a comparison of the BISs in terms of these characteristics (de Luis-Garcia et al., 2003). Among the technologies in the table, fingerprint technology accounts for the greatest share of the global market (de Luis-Garcia et al., 2003; BCC, 2010) despite its low user-acceptance and medium cost.

Table 1.1. Characteristics of biometric identification systems (de Luis-Garcia et al., 2003).

	Accuracy	Ease of use	User-acceptance	Ease of implementation	Cost
Fingerprint	High	Medium	Low	High	Medium
Face	Low	High	High	Medium	Low
Iris	Medium	Medium	Medium	Medium	High
Retina	High	Low	Low	Low	Medium
Hand geometry	Medium	High	Medium	Medium	High
Signature	Medium	Medium	High	Low	Medium
Voice	Medium	High	High	High	Low

When Table 1.1 is examined, it is observed that voice is one of the most promising biometric technologies since its user acceptance, ease of use and ease of implementation are high and its cost is low. Additionally, it is the most natural biometric information source in today's telephone based remote access control applications. In the table, the only medium rate for voice is its accuracy. There are various reasons which may degrade the performance of a voice biometric or speaker recognition system. First, although each person has a unique vocal tract, voice is inherently more vulnerable to mimicry when compared to fingerprint. Second, voice is greatly affected by the amount of stress and other factors such as aging and illness. Third, high background noise and varying recording conditions significantly reduce the accuracy of speaker recognition systems. Before starting to discuss these issues in more detail, we will first define the problem and introduce some terminology in the next section.

1.2. Problem Definition

Speaker recognition is the task of recognizing a person from his/her voice. Speaker recognition applications can be divided into two major categories: identification and verification. In speaker identification, we try to determine which speaker out of a group of known speakers produces the input voice sample. There are two modes of operation: closed-set and open-set. In closed-set identification, the most likely registered speaker is chosen as the identity of the test sample. It is a multiple class classification problem, and the number of classes equals to the number of speakers, *N*, in the population. When the process also includes declaring a speech sample when it does not belong to any of the registered speakers, then it is referred to as open-set identification (Ariyaeeinia et al., 2006). There are $N+1$ decision classes in open-set problem. The goal of speaker verification is to verify the claimed identity of an unknown speaker. This task is also known as speaker authentication, voice verification and voice authentication. It can be considered as a true-or-false binary decision problem.

A speaker recognition system typically involves two phases: enrollment and authentication. During enrollment, a user provides voice samples to the system. The system extracts speaker-specific features from the voice samples to build a voice model of the user. The voice model is also called as speaker model (or voiceprint). In authentication phase, a test voice sample is provided from an unknown user and it is compared to the enrolled speaker models for an identification or verification decision. The speaker associated with the model that is being tested is referred to as target speaker or claimant (Jin, 2007). In speaker verification, the test sample is only scored against the claimant model. Identity claim of the user is accepted (or rejected) based on the similarity score. In closed-set speaker identification, test sample is compared to all models in the system and the best matching model is chosen as the identity of the user. Open-set identification systems should also be able to reject the test samples that do not belong to any of the registered speakers in addition to the identification decision for known speakers. Therefore, it can be viewed as the merger of closed-set identification and verification tasks (Reynolds and Campbell, 2007).

Speaker recognition can be further classified into text-independent and text-dependent categories according to the constraints placed on the enrollment and authentication speech (Reynolds and Campbell, 2007; Jin, 2007). In a text-independent system, there is no constraint for the speech samples. Therefore, these systems offer more flexible scenarios, e.g., allowing verification of a speaker while he/she speaks for some other purpose. In a text-dependent task, user must speak a given phrase known to the system. The phrase can be common for all speakers or unique. It can be a fixed text or can be randomly generated at the test time from a limited vocabulary (such as digits). Although text-dependent systems require user cooperation, they can provide better recognition accuracy for relatively short enrollment and authentication utterances due to the prior knowledge of the constrained text.

Speaker recognition systems can also be distinguished by the constraints imposed by the target application. Depending on the application type, the speech can be collected from a noise-free environment with a wide-band microphone or from a noisy environment using a narrow-band telephone channel. In telephone-based applications, robustness to channel variations might be a major concern. The data amount may also range from several seconds to several hours. It is common to expect that the verification performance will increase with increasing constraints on the specific application (e.g., more speech, less noise, known channels) (Reynolds, 2002).

1.3. A Very Brief Overview

The first steps of speaker recognition can be dated back to the development of sound spectrograph in the Bell telephone laboratories (Koenig et al., 1946). The sound spectrograph is an instrument used to analyze complex speech sounds and their variations in time. The analysis is accomplished with the use of spectrogram representation. Then, Lawrence Kersta was the first person who suggested the visual comparison of the speech spectrograms for speaker identification (Kersta, 1962). Due to high recognition rate in Kersta's work and claim that the accuracy of voiceprints might be comparable to that of fingerprints in the forensic setting (Poza and Begault, 2005), spectrogram based techniques have continued to receive more attention (Bolt et al., 1970; Tosi et al., 1972). Long-term averaged spectrum of a sentence-long utterance was also used for speaker identification in early studies (Furui, et al., 1972; Markel et al., 1977; Furui, 2009). Additionally, a significant effort has been made to reduce the effects of mismatch channel conditions since the main reason of performance degradation in telephone based applications has been attributed to this mismatch.

Coming to more recent days, text-independent NIST evaluations (NIST, 2011) have contributed significantly to the development and calibration of new technologies. Modern speaker recognition systems consist of several individual subsystems which are combined for the best possible performance (Sturim et al., 2009; Kajarekar et al., 2009; Li et al., 2009). Gaussian mixture models (GMMs) and support vector machines (SVMs) are two popular classification methods. High-level features such as prosody, duration and word frequencies have provided complementary information to the low-level spectral features. Factor analysis (FA) (Kenny, 2005) in GMM-based and nuisance attribute projection (NAP) (Solomonoff et al., 2005) in SVM-based systems are the state-of-art channel compensation techniques for text-independent tasks.

Although text-independent studies have dominated the recent years, researches have continued for text-dependent tasks. In this paragraph, we make a brief review of recent text-dependent studies. Some of the studies will be discussed in more detail in the following sections. In (Hebert and Boies, 2005), effect of lexical mismatch between target and cohort speaker's lexicons is investigated for test normalization (T-norm) which is a frequently used score normalization technique especially for text-independent tasks. In (Toledano et al., 2008), the same lexical mismatch problem is addressed using phoneme and sub-phoneme level T-norm. In (Li et al., 2007), a procedure to combine T-norm and cohort normalization is presented. In (BenZeghiba and Bourlard, 2006), use of multiple background models for likelihood normalization and multiple reference models for speaker model training is investigated. The relative importance of temporal characteristics is studied in (Nealand et al., 2005) by comparing the performances of various hidden Markov model (HMM) and GMM configurations. In (Ramasubramanian, 2006), multiple templates are used for each word to implement a variable text text-dependent speaker recognition system using a dynamic programming algorithm. In (Das et al., 2008), a new database is presented for text-dependent tasks. In the study, multi-template dynamic time warping (DTW), HMM and text-conditioned vector quantization (VQ) methods are compared using the self-collected database. Direct modeling of the spoken password by a fixed dimensional feature vector is explored in (Das and Tapaswi, 2010). As a result, overall storage and computational requirements are reduced with the proposed method when compared to DTW and HMM methods. SVM based text-dependent speaker verification using HMM supervectors is studied in (Dong et al., 2008). Although SVM based method did not perform better than the traditional GMM and HMM based methods, score level fusion of the methods lead to improvement in verification performance. A discriminative algorithm is used in conjunction with a generative model (HMM) in (Subramanya et al, 2007). In the method, speakers are modeled with the generative model and the discriminative model is used for decision making. In (Yegnanarayana, 2005), source and suprasegmental (pitch and duration) features are combined with a baseline spectral system in a single utterance task. In (Mirghafori and Hebert, 2004), a strategy is presented for setting a priori threshold in an adaptive text-dependent speaker verification system. The threshold is calculated as a function of length of the password, the number of training frames in the speaker model and target false acceptance rate.

1.4. Work Done

Over the recent years, much of the effort in speaker recognition community has been concentrated on text-independent applications. This can be mainly attributed to almost annual NIST evaluations. However, text-dependent speaker verification has gained more attention in private sector for fraud prevention because of ease of use and higher accuracy for relatively short enrollment and authentication utterances. Those systems also offer significant cost reduction in call centers since they can reduce or eliminate the need for identity check questions. In this thesis, we investigate model selection and channel variability issues for two text-dependent speaker verification applications.

For years, YOHO (Higgins et al., 1991; Campbell, 1995) has been better known for evaluation of text-dependent speaker recognition. The database consists of "combination lock" phrases (e.g. "35 - 72 – 41") in English. However it lacks of channel variations since the recordings are taken with a high-quality telephone handset and they are not passed through a telephone channel (Campbell, 1995). To the best our knowledge, there is no commercially available Turkish database to study handset-channel variability issues for text-dependent tasks. Therefore, we design our own multi-channel databases. The databases are named as text-dependent variable text database (TDVT-D) and text-dependent single utterance database (TDSU-D). TDVT-D includes 52 speakers over 5 different handset-channel conditions and consists of Turkish digit sequences and short phrases. TDSU-D includes 59 speakers over 5 different handset-channel conditions and contains a single Turkish phrase.

Using the databases, we devote the first set of experiments to model selection. GMM has been the dominant classification method especially for text-independent speaker verification (Reynolds et al., 2000). However, HMMs can be an appropriate choice for text-dependent tasks since they can capture co-articulation information better. In TDVT-D, we compare the performances of GMM based and context independent HMM based methods for several different authentication utterances, speaker model adaptation scenarios and enrollment-authentication channel conditions. In the experiments, we make use of a rank-based decision-making procedure. In the procedure, each authentication utterance is scored against a number of cohort speaker models in addition to the claimant model. Then the scores are sorted in descending order. We use the rank of claimant to make the verification decision. In TDSU-D, GMM and two HMM based techniques are compared to find the most appropriate classification method for the TDSU task. In the first HMM-based approach, 3-state context independent HMMs are trained for each phoneme in the utterance. In the second approach, a single whole-phrase sentence HMM is matched to the fixed utterance to better capture co-articulation information.

In our model selection experiments, we observe a significant performance gap between match and mismatch condition results and thus devote the second set of experiments to channel mismatch compensation. In speaker verification literature, various compensation techniques are proposed in order to reduce the effects of the channel mismatch. The techniques are mainly applied in three domains; feature domain, model domain and score domain. In TDVT-D, we investigate three compensation techniques in the three domains. These techniques are named as long time average spectrum (LTAS) filtering, maximum likelihood linear regression (MLLR) transformation and handset-dependent rank-based decision making (H-rank) which are applied in feature, model and score domain, respectively. In LTAS filtering during the authentication, short time spectrum of speech signal is filtered by a pre-computed LTAS filter which is estimated for each enrollment and authentication channel pair. In MLLR transformation, claimant speaker's model is adapted to the authentication channel condition using a pre-computed MLLR matrix. In H-rank, we benefit from the prior knowledge of enrollment handset-channel type in order to perform the ranking among the cohorts who share the same handset with the claimant's enrollment.

In TDSU channel compensation experiments, we first concentrate on score domain methods. We investigate two popular score normalization techniques for the TDSU task, namely test normalization (T-norm) and zero normalization (Z-norm), in addition to the rank-based decision making. In recent NIST evaluations, combination of the normalization methods improved text-independent verification performance (Aronowitz et al., 2005; Kenny et al., 2007; Vogt et al., 2005). Therefore, we implement two possible combinations (i.e., TZ-norm and ZT-norm) and propose a novel combination procedure referred to as C-norm. Handset-dependent versions of each method are also studied in this section. Next, we present a feature domain compensation method referred to as spectral division (SMD) as an enhancement to the well-known cepstral mean subtraction (CMS) or normalization (CMN) and compare the performances of the two methods. In the last section, we investigate several prosodic features for the TDSU task and combine them with the baseline cepstral features (we will refer them as spectral features) to further improve the verification accuracy.

In most of the aforementioned compensation methods, a channel recognition system is required in order to correctly apply the appropriate normalization parameters to the features, models or scores. A GMM structure is commonly used for this purpose (Reynolds et al., 2000; Mak et al., 2002; Mak et al., 2004). Throughout the channel compensation experiments, we also address the effects of channel detection errors on speaker verification performance. Moreover, we present a cohort-based channel detection procedure in addition to the GMM based method. In the cohort-based method, some imposter speaker likelihoods from several different channel conditions are used to identify the employed channel type in enrollment and authentication.

1.5. Contributions

The contributions of this thesis can be summarized with the following items:

- We collected two text-dependent Turkish speaker recognition databases. In the future, we plan to distribute the databases for academic research purposes.
- Although GMM and HMM based method are extensively studied in the literature, we make a comprehensive comparison and evaluation of the methods for several different authentication utterances, enrollment scenarios and enrollment-authentication channel conditions.
- We investigate MLLR transformation in model domain, LTAS filtering and SMD in feature domain for channel compensation in HMM-based text-dependent framework.
- We use a rank-based procedure for decision making and compare its performance with other well-known score normalization methods such as T-norm and Z-norm.
- We compare various score normalization methods for the TDSU task and propose a novel normalization method referred to as C-norm.
- We extensively study the effects of employing channel detection prior to channel compensation. We propose a cohort-based channel detection procedure in addition to the classical GMM based method.

- We investigate several prosodic features (pitch, duration and energy) for the TDSU task and combine them with the baseline spectral system using time alignment of HMM states.

1.6. Outline of thesis

The remainder of this thesis is organized as follows. In Chapter 2, we provide an overview of speaker recognition theory. In Chapter 3, channel mismatch compensation techniques are introduced. The databases are presented in Chapter 4. Chapter 5 and Chapter 6 are devoted to experimental results for TDVT and TDSU databases, respectively. In Chapter 7, we conclude the thesis with a summary of results and observations.

2. THEORY OF SPEAKER RECOGNITION

Our main focus in this thesis is the speaker verification problem. Therefore, we will make a theoretical review of this problem in this chapter. Only in decision-making section, we will briefly mention open-set and closed-set identification tasks. Given a test speech segment, *Y*, and a claimant speaker, *S*, the task of speaker verification is to determine if *Y* was spoken by *S*. Since we assume that *Y* contains speech from only one speaker, the task is better termed single-speaker verification (Reynolds et al., 2000).

The single-speaker verification problem can be stated as a basic hypothesis test between null (H_0) and alternative (H_1) hypotheses,

H_0 = *Y* is from the claimed speaker *S*.
H_1 = *Y* is not from the claimed speaker *S*.

A likelihood ratio test is used to decide between the hypotheses

$$\frac{P(Y|H_0)}{P(Y|H_1)} \geq \theta \; accept \; H_0 \tag{2.1}$$

There are two types of errors in speaker verification systems. In the first error type (Type I), the null hypothesis is incorrectly rejected when it is true. This error type is referred to as miss or false rejection. In Type II error, null hypothesis is incorrectly accepted when the alternative is true. This error type is known as false alarm and false acceptance. The cost of each error type depends on the target application. For example in a secure telephone banking application with voice authentication, accepting an imposter speaker might produce more fatal consequences than rejecting a true client.

In Figure 2.1 and Figure 2.2, enrollment and authentication stages of a speaker verification system are shown. The blocks in the figures will be discussed in the following sections. In Section 2.1, front-end processing block will be presented. Section 2.2 is devoted to speaker and background models. Decision making component will be described in Section 2.3.

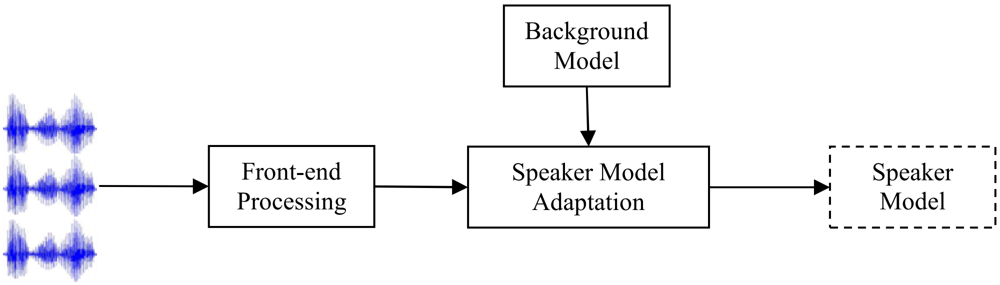

Figure 2.1. Enrollment stage of a speaker verification system.

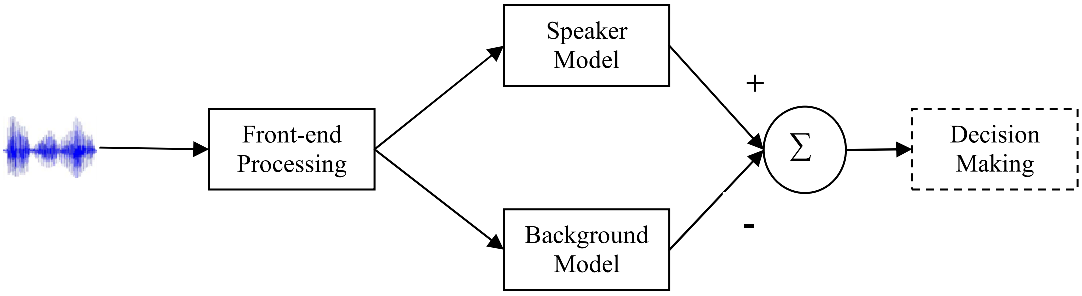

Figure 2.2. Authentication stage of a speaker verification system.

2.1. Front-end Processing

The role of front-end processing block is to extract features from the speech signal that convey speaker-specific information. Other pre-processing steps such as end-point detection and noise removal are also employed at this stage of the system. We can classify the features used in speaker recognition applications into two main categories: short-term (lower-level) and long-term (higher-level) features. Short-term features are computed from a short frame of speech and they are related to the anatomical structure of the speaker's vocal tract. On the other hand, long-term features span longer time intervals and capture phonetic, prosodic and lexical information.

2.1.1. Short-term features

Today most speaker recognition systems rely on the short-term spectral features extracted from a short segment of speech signal. Mel-frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980), linear prediction cepstral coefficients (LPCCs) (Makhoul, 1975), perceptual linear predictive (PLP) coefficients (Hermansky, 1990) and line spectral frequencies (LSFs) (Itakura, 1975) are commonly used in speech and speaker recognition applications. These features mainly differ in the details of spectrum representation. Since MFCCs are very popular in speaker recognition tasks, we will also use this feature in this thesis.

A filter-bank analysis is performed to compute MFCCs as shown in Figure 2.3.

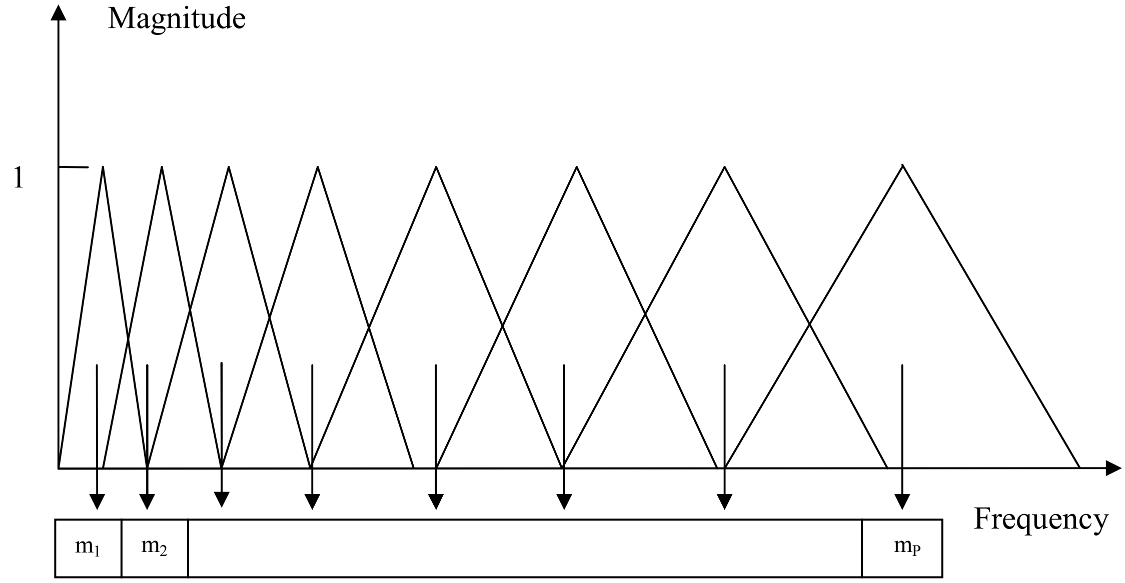

Figure 2.3. Mel-scale filter bank.

In the analysis, speech signal is divided into overlapping frames of 20-30 msec window length and 10-15 msec frame shift. First, discrete Fourier transform (DFT) of the windowed speech segment is taken and its magnitude is computed. In this step, we obtain DFT magnitude of the speech frame, X_k where k is the DFT index. Then, X_k is multiplied by the corresponding triangular filter gain in Figure 2.3 and the results are accumulated. As a result, each filter-bank coefficient (m_i) represents a weighted sum of spectral magnitude in that filter-bank channel.

$$m_i = \sum_k w_{k,i} X_k \tag{2.2}$$

where $w_{k,i}$ represent the weights of the filter (they are assumed to be zero outside the band of the filter).

In MFCC computation, the filter-banks are not equally distributed across the acoustic spectrum (*f*) but they are equally spaced along the mel-scale which is defined by

$$Mel(f) = 2595 * \log\left(1 + \frac{f}{700}\right) \tag{2.3}$$

This is due to the fact that the human ear resolves frequencies non-linearly across the spectrum and it is empirically observed that designing a front-end to operate in a similar non-linear manner improves recognition accuracy.

Finally, discrete cosine transform (DCT) of the log filter-bank magnitudes are computed to obtain the final coefficients as follows:

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} \log\,(m_j) \cos\left(\frac{\pi i}{N}(j - 0.5)\right) \tag{2.4}$$

where *N* is the number of filter-bank channels. Only first few coefficients of the DCT are kept in the feature vector and they are referred to as static coefficients. In general, time derivatives of the static coefficients are appended to the feature vector to enhance the performance of recognition systems. First order regression coefficients (also referred to as delta coefficients) are calculated using the following formula

$$\boldsymbol{\delta}_t = \frac{\sum_{k=1}^{\theta} k(\boldsymbol{c}_{t+k} - \boldsymbol{c}_{t-k})}{2\sum_{k=1}^{\theta} k^2} \tag{2.5}$$

where $\boldsymbol{\delta}_t$ is the delta coefficient at time *t* which is computed using the static coefficients from time *t-k* to *t+k*.

At the end of front-end processing block, the speech signal is represented with a sequence of feature (or observation) vectors, $\boldsymbol{O} = [\boldsymbol{o}_1, \boldsymbol{o}_2, \boldsymbol{o}_3 \ldots \boldsymbol{o}_T]$ where the subscript stands for the frame or time index.

2.1.2. Long-term features

In speaker recognition literature, there has been a significant effort to utilize higher-level information sources for more accurate and robust speaker recognition and much of this effort has been concentrated on text-independent applications. To better analyze the high-level features, we can divide them into three types; phonetic, lexical and prosodic. Time sequence of phones and speaker-dependent pronunciation modeling are among the phonetic features. They are generally extracted using the outputs of a speech recognizer in text-independent applications. Lexical features are also obtained with the help of a speech recognizer and include speaker-dependent word frequencies. Prosodic features aim to capture variation in intonation, timing and loudness. Pitch, duration and energy are the most studied prosodic features.

Generally, use of high-level features requires more training data compared to the low-level spectral features. Moreover, in order to extract some of them, a complex and accurate speech recognition system is needed. However, with the availability of highly accurate speech recognizers and relatively large amount of enrollment data from a speaker, they have provided complementary information to the spectral features in text-independent applications (Weber et al., 2002; Reynolds et al., 2003; Klusacek et al., 2003; Adami et al., 2003; Shriberg et al., 2005; Dehak et al., 2007, Ferrer et al., 2010). Moreover, they are known to be less susceptible to channel variations and thus may further improve the verification performance under mismatch channel conditions.

For text-dependent applications, some of the high-level features (e.g., time sequence of phones and speaker-dependent word frequencies) are not meaningful. Additionally, enrollment data amount in commercial text-dependent applications may not be adequate to reliably estimate the parameters of some features (e.g., speaker-specific pronunciations). However, especially in TDSU task, prosodic features may provide complementary information to the spectral features. In this thesis, we study pitch, duration and energy features for our TDSU task and combine them with the baseline spectral system for more accurate and robust verification.

2.2. Modeling Approaches

Modeling approaches used in speaker verification applications can be classified into two main categories, generative models and discriminative models. In generative modeling, speakers are explicitly represented with probability density functions. Vector quantization (VQ), Gaussian mixture models (GMMs), hidden Markov models (HMMs) and dynamic time warping (DTW) are examples of generative models. On the other hand, discriminative models such as support vector machines (SVMs) and artificial neural networks (ANNs) estimate a boundary between target and imposter speakers.

SVMs have been a very popular tool recently especially after speaker recognition community has re-discovered a robust way to represent utterances (Kinnunen and Li, 2010) with a single vector which is referred to as supervector. SVMs also lead to improvement in verification performance when combined with generative models in text-independent (Campbell et al., 2006) and text-dependent (Dong et al., 2008) applications. However SVM modeling of HMM supervectors did not perform better than HMM and GMM baseline systems in a fixed password task (Dong et al., 2008).

Despite the recent success of discriminative models especially in text-independent applications, our focus will be on the generative models in this thesis. VQ is one of the simplest text-independent generative models. In the enrollment stage of VQ, the training vectors are clustered into K reference vectors, $\boldsymbol{R} = [\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3 ... \boldsymbol{r}_K]$, using a clustering algorithm such as K-means (Linde et al., 1980). Then, the average quantization distortion between the reference and test utterance feature vectors is computed for decision-making.

$$D = \frac{1}{T}\sum_{t=1}^{T} \min_{1\leq k\leq K} d(\boldsymbol{o}_t, \boldsymbol{r}_k) \tag{2.6}$$

where d(·,·) is a distance measure (e.g., the Euclidian distance). GMM is another famous generative model and has been the dominant modeling technique especially for text-independent applications. It has also become a standard reference method in speaker recognition. Text information is also incorporated to GMMs for possible performance improvement in (Gutman and Bistritz, 2002; Sturim et al., 2002). Additionally, GMM supervectors with joint factor analysis (JFA) compensation have been one of the most successful systems in recent NIST evaluations (Kinnunen and Li, 2010).

DTW is a generative model especially used in text-dependent applications. It is a cost minimization matching algorithm that accomplishes the time alignment of test and reference features through a dynamic programming procedure. During the enrollment, a single reference template is obtained by first aligning the enrollment utterances with each other using DTW and then averaging the aligned features (Furui, 1981; Yu et al., 1995). DTW is inherently text-dependent and it is difficult to implement a text-prompted system with it although it is not completely impossible (Ramasubramanian et al., 2006). Although HMMs are more complex compared to DTW, they provide greater flexibility. For example, we can implement a text-independent application with HMMs by running a word or phone level speech recognizer (Newman et al., 1996; Lamel and Gauvain, 2000). HMM can also be considered as the generalization of GMM where the states of the model constraint the alignment of feature vectors to appropriate mixture component. Thanks to this property, they might be a good choice for text-dependent applications. With respect to the target application, whole-sentence (for TDSU tasks), word, or sub-word HMMs might be utilized. Continuous/semi-continuous, left-to-right/ergodic HMM topologies might also preferred.

Especially in the late 1990s, most of the aforementioned generative models are compared for various text-independent and text-dependent tasks (Matsui and Furui, 1993; Matsui and Furui, 1994a,b; Zhu et al., 1994; Reynolds and Rose, 1995; Yu et al., 1995;

Falavigna, 1995; van Vuuren, 1996). In our thesis, we also compare different HMM topologies with a reference GMM based method using our TDVT and TDSU databases.

In GMM and HMM methods, both null and alternative hypotheses are governed by probability density functions. The null hypothesis is represented with a speaker model, Λ_S, which characterizes the claimant speaker in acoustic feature space. The alternative hypothesis is governed by a background model, λ_{BKG}, which should represent all possible alternatives to the target speaker and cover the broad acoustic classes of speech sounds. In Section 2.2.1, we first present theory of GMM and HMM based methods. Then, we discuss the background model in Section 2.2.2 and a way of obtaining speaker models by adapting the parameters of background model in Section 2.2.3.

After defining the speaker and background models, the likelihood ratio test in Equation 2.1 can be written as:

$$\Lambda_S(\boldsymbol{O}) = \log P(\boldsymbol{O}|\lambda_S) - \log P(\boldsymbol{O}|\lambda_{BKG}) \tag{2.7}$$

As seen in the equation, generally logarithm of the likelihood ratio statistic is used in speaker recognition. If we assume that the observation vectors, $\boldsymbol{o}_t$, are independent, we can compute the log likelihood ratio as follows:

$$\Lambda_S(\boldsymbol{O}) = \sum_{t=1}^{T} (\log P(\boldsymbol{o}_t|\lambda_s) \; - \; \log P(\boldsymbol{o}_t|\lambda_{BKG})) \tag{2.8}$$

2.2.1. GMM and HMM based methods

2.2.1.1. GMM-based Method. GMM has been the most widely used modeling technique especially for text-independent applications since it is computationally inexpensive, is based on a well-understood statistical model and is insensitive to the temporal aspects of the speech. However, for text-dependent applications, the last advantage of the GMM might be a disadvantage since temporal characteristics might convey some information about the speaker identity.

In GMM, probability density functions for the background and speaker models are mixtures of Gaussians;

$$P(\boldsymbol{o}_t|\lambda) = \sum_{i=1}^{M} w_i N(\boldsymbol{o}_t, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{2.9}$$

where M is the number of mixtures and w_i is the mixture weight for i^{th} mixture. The mixture weights satisfy the following unity constraint:

$$\sum_{i=1}^{M} w_i = 1 \tag{2.10}$$

In Equation 2.9, each mixture component is governed by a normal (Gaussian) distribution, $N(:, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, which is parameterized by a *D x 1* mean vector, $\boldsymbol{\mu}_i$, and a *D x D* covariance matrix, $\boldsymbol{\Sigma}_i$, where *D* is the dimension of the observation vectors.

$$N(\boldsymbol{o}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} * exp\left[-\frac{1}{2}(\boldsymbol{o} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i{}^{-1}(\boldsymbol{o} - \boldsymbol{\mu}_i)\right] \tag{2.11}$$

Complete parameter set of a GMM is denoted with the following notation

$$\lambda = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{\mathrm{i}}\} \qquad i = 1, \dots, M \tag{2.12}$$

Although the general form of the modeling supports the use of full covariance matrices, only diagonal entries of the matrix are used in practice since diagonal covariance GMMs are computationally more efficient and they empirically outperform full covariance GMMs (Reynolds et al., 2000). Therefore we can describe a GMM with $\lambda = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\sigma}_{\mathrm{i}}\}$ where $\boldsymbol{\sigma}_i^2 = \boldsymbol{\Sigma}_{ii}$.

2.2.1.2. HMM-based Method. A hidden Markov model can be considered a generalization of a GMM where hidden states control the mixture component to be selected for each observation vector. In HMM, it is assumed that the sequence of observed speech vectors is generated by a Markov model. A Markov model is a finite state machine which changes state once every time unit and at each time *t* that a state *j* is entered, an observation vector ($\boldsymbol{o}_t$) is generated by the probability density $b_j(\boldsymbol{o}_t)$. Generally, each state is modeled by a mixture of Gaussians. If the observation vector, $\boldsymbol{o}_t$, is aligned to state *j*, then $b_j(\boldsymbol{o}_t)$ can be written as

$$b_j(\boldsymbol{o}_t) = \sum_{k=1}^{M} w_{jk} N(\boldsymbol{o}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) \tag{2.13}$$

where M is the number of mixtures, w_{jk} is the mixture weight for k^{th} mixture and $N(\boldsymbol{o}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk})$ is the normal distribution. Mixture weights in the equation satisfy the unity constraint.

In the modeling, transition from state *i* to state *j* is also probabilistic and is governed by the discrete transition probability a_{ij} which satisfies the following constraint

$$\sum_{j=1}^{N} a_{ij} = 1 \quad \forall i \tag{2.14}$$

where *N* is the number of states in HMM.

Complete parameter set of an HMM is $\lambda = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\}$ where $\boldsymbol{\pi} = \{\pi_i\}$ represents initial state probabilities

$$\pi_i = P(q_1 = \mathrm{i}) \qquad 1 \leq i \leq N \tag{2.15}$$

In the equation, q_t denotes the state at time *t* and thus q_1 is the initial state.

Given a state sequence $\boldsymbol{q} = [q_1, q_2, q_3 ... q_T]$ probability of the observation vectors, $\boldsymbol{O} = [\boldsymbol{o}_1, \boldsymbol{o}_2, \boldsymbol{o}_3 ... \boldsymbol{o}_T]$ can be written as in Equation 2.16.

$$P(\boldsymbol{O}|\boldsymbol{q},\lambda) = \prod_{t=1}^{T} P(\boldsymbol{o}_t|q_t,\lambda) = b_{q_1}(\boldsymbol{o}_1)\, b_{q_2}(\boldsymbol{o}_2) \dots b_{q_T}(\boldsymbol{o}_T) \tag{2.16}$$

In the equation, we assume that the observation vectors are independent. We can write probability of such a state sequence $\boldsymbol{q}$ as follows:

$$P(\boldsymbol{q}|\lambda) = \pi_{q_1} a_{q_1 q_2} \dots a_{q_{T-1} q_T} \tag{2.17}$$

The joint probability of the observation vectors and state sequence is a simple product of Equations 2.16 and 2.17.

$$P(\boldsymbol{O},\boldsymbol{q}|\lambda) = P(\boldsymbol{O}|\boldsymbol{q},\lambda)P(\boldsymbol{q}|\lambda) \tag{2.18}$$

The probability of $\boldsymbol{O}$ (given the model) is obtained by summing the joint probability over all possible state sequences.

$$P(\boldsymbol{O}|\lambda) = \sum_{\forall\, Q} P(\boldsymbol{O}|\boldsymbol{q},\lambda)P(\boldsymbol{q}|\lambda) \tag{2.19}$$

Although the calculation of $P(\boldsymbol{O}|\lambda)$ from its direct definition in Equation 2.19 is computationally infeasible, simple recursive procedures exist to efficiently compute it. For this purpose, Viterbi algorithm is generally used when the calculation is based on the most likely state sequence. To find the best state sequence $\boldsymbol{q} = [q_1, q_2, q_3 ... q_T]$ given the observation vectors, we define the quantity

$$\delta_t(i) = \max_{q_1, q_2, \dots q_{t-1}} P(q_1 q_2 ... q_{t-1}, q_t = i, \boldsymbol{o}_1 \boldsymbol{o}_2 ... \boldsymbol{o}_t|\lambda) \tag{2.20}$$

that is the best score along a single path, at time *t*, which accounts for the first *t* observations and ends in state *i*. By induction we have

$$\delta_{t+1}(j) = \left[\max_{i} \delta_t(i) a_{ij}\right] b_j(\boldsymbol{o}_{t+1}) \tag{2.21}$$

To find the best state sequence, we need to keep track of the argument that maximizes Equation 2.21 for each time *t* and state *j*. This is done via the array, $\psi_t(j)$. After defining the required quantities, Viterbi algorithm can be written in four steps.

(i) Initialization

$$\delta_1(i) = \pi_i b_i(\boldsymbol{o}_1) \qquad 1 \le i \le N \tag{2.22}$$

$$\psi_1(i) = 0 \tag{2.23}$$

(ii) Recursion

$$\delta_t(j) = \max_{1 \le i \le N}\left[\delta_{t-1}(i) a_{ij}\right] b_j(\boldsymbol{o}_{t+1}) \quad 2 \le t \le T, 1 \le j \le N \tag{2.24}$$

$$\psi_t(j) = \underset{1 \le i \le N}{\operatorname{argmax}}\left[\delta_{t-1}(i) a_{ij}\right] \quad 2 \le t \le T, 1 \le j \le N \tag{2.25}$$

(iii) Termination

$$p^* = \max_{1 \le i \le N}[\delta_T(i)] \tag{2.26}$$

$$q_T^* = \underset{1 \le i \le N}{\operatorname{argmax}}[\delta_T(i)] \tag{2.27}$$

(iv) State sequence backtracking

$$q_t^* = \psi_{t+1}(q_{t+1}^*) \qquad t = T-1, T-2, \dots, 1 \tag{2.28}$$

In HMM-based methods, we obtain the speaker scores by forced alignment of the observation vectors to the known text using Viterbi algorithm. This alignment procedure is illustrated in Figure 2.4. As observed in the figure and pointed out in this subsection, HMM states constrain the mixture component to be selected for each observation vector.

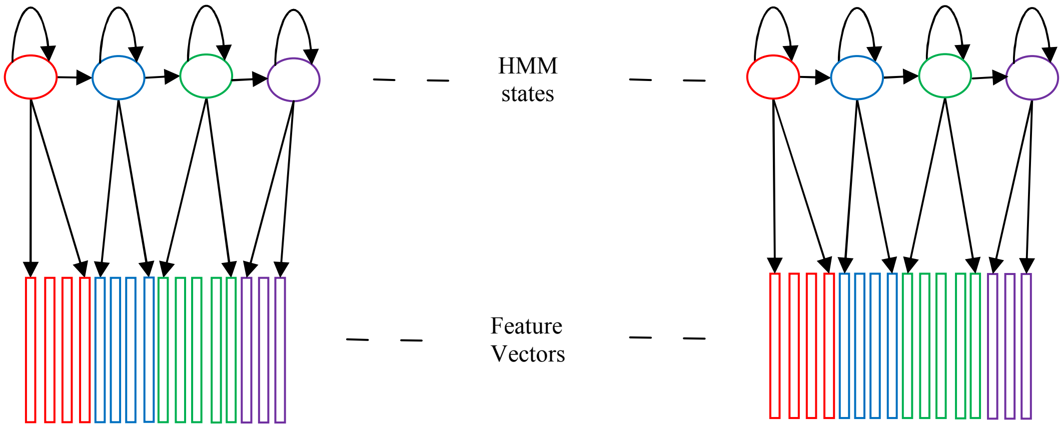

Figure 2.4. Alignment of feature vectors to HMM states.

2.2.2. Background Model

There are two well-known alternative approaches for the background modeling. In this subsection, we will present these two alternatives and how $P(\boldsymbol{O}|\lambda_{BKG})$ can be computed in the alternatives. In the first approach, a set of imposter speaker models are used to represent the background model. These imposters are also named as cohort speakers. Given a set of *N* cohort speakers, the alternative hypothesis is represented by the average (or maximum) of the cohort speaker likelihoods as follows:

$$P(\boldsymbol{O}|\lambda_{BKG}) = \frac{1}{K}\sum_{k=1}^{K} P(\boldsymbol{O}|\lambda_k) \tag{2.29}$$

where K denotes the number of cohort speakers and λ_k represents the model of k^{th} cohort speaker. The size and selection of the cohort speakers plays an important role in this approach. A common set for all test speakers might be preferable since it is easy to construct and does not require many cohorts. However, a unique set for each target speaker might improve the performance when constructed using some kind of speaker similarity measure (Rosenberg and Parthasarathy, 1996; Isobe and Takahashi, 1999). On the other hand, the use of speaker-specific cohorts might be a drawback for some applications which use a large number of target speakers (Reynolds et al, 2000).

In the second approach, speech from various speakers is pooled to train a single speaker independent model. This speaker independent model is also referred to as general model or universal background model (UBM). This approach has an advantage compared to the first approach, since a single background model can be trained once and used for all target speakers. On the other hand, some performance improvement can be obtained with the use of handset and gender dependent background models (Heck and Weintraub, 1997).

When the second approach is taken for the alternative hypothesis modeling, the parameters of the background model should be trained using a speech database. We will now describe the most popular and well-established method for the parameter estimation assuming a GMM configuration (Reynolds, 2008). Given a collection of training vectors, $\boldsymbol{O} = [\boldsymbol{o}_1, \boldsymbol{o}_2, \boldsymbol{o}_3 ... \boldsymbol{o}_T]$, maximum likelihood estimates of the model parameters are computed via the iterations of expectation-maximization (EM) algorithm. In each EM step, the likelihood of the estimated model for the feature vectors is monotonically increased.

$$P(\boldsymbol{O}|\lambda^{k+1}) > P(\boldsymbol{O}|\lambda^{k}) \tag{2.30}$$

where k denotes the iteration number. $P(\boldsymbol{O}|\lambda)$ is defined with the assumption that the training vectors are independent of each other.

$$P(\boldsymbol{O}|\lambda) = \prod_{t=1}^{T} P(\boldsymbol{o}_t|\lambda) \tag{2.31}$$

If the model parameters are refined using the re-estimation formulas in Equation 2.32-2.34, a monotonic increase in the model's likelihood is guaranteed.

$$w_i^{k+1} = \frac{1}{T}\sum_{t=1}^{T} P(i|\boldsymbol{o}_t, \lambda^k) \tag{2.32}$$

$$\boldsymbol{\mu}_i^{k+1} = \frac{\sum_{t=1}^{T} P(i|\boldsymbol{o}_t, \lambda^k)\boldsymbol{o}_t}{\sum_{t=1}^{T} P(i|\boldsymbol{o}_t, \lambda^k)} \tag{2.33}$$

$$\boldsymbol{\Sigma}_i^{k+1} = \frac{\sum_{t=1}^{T} P(i|\boldsymbol{o}_t, \lambda^k)(\boldsymbol{o}_t - \boldsymbol{\mu}_i^{k+1})(\boldsymbol{o}_t - \boldsymbol{\mu}_i^{k+1})'}{\sum_{t=1}^{T} P(i|\boldsymbol{o}_t, \lambda^k)} \tag{2.34}$$

where the prime stands for vector or matrix transpose. $P(i|\boldsymbol{o}_t, \lambda^k)$ is the probabilistic alignment of the training vector, $\boldsymbol{o}_t$, to i^{th} mixture component of the model, $\lambda^k = \{w_i^k, \boldsymbol{\mu}_i^k, \boldsymbol{\Sigma}_i^k\}$ and it is computed as follows:

$$P(i|\boldsymbol{o}_t) = \frac{w_i^k N(\boldsymbol{o}_t, \boldsymbol{\mu}_i^k, \boldsymbol{\Sigma}_i^k)}{\sum_{j=1}^{M} w_j^k N(\boldsymbol{o}_t, \boldsymbol{\mu}_j^k, \boldsymbol{\Sigma}_j^k)} \tag{2.35}$$

where M is the number of mixtures. The iterations of EM algorithm are repeated until a convergence threshold is reached. EM will find a local maximum of the likelihood function, and this local solution may not correspond to the global solution due to the starting point of the algorithm (i.e., λ^0). Therefore, initialization of the model parameters might have a critical role to arrive at a good final point with minimum number of iterations. However, experimental results showed that elaborate initialization schemes are not necessary in speaker verification applications and random initial means and identity initial covariance matrices perform comparably to the other more complicated initialization procedures (Reynolds et al., 1995).

2.2.3. Speaker Model

Speaker models might be trained from scratch similar to the background model using the speaker's enrollment speech. However, the enrollment data amount may not be adequate to reliably estimate the large number of model parameters in this case. A second alternative is to adapt the well trained parameters of the background model to the feature space of the target speaker. Various speaker model adaptation techniques are studied in speaker and speech recognition literature (Mokbel, 2001; Mak et al., 2006) and maximum likelihood linear regression (MLLR) transformation and maximum a posterior (MAP) adaptation are among the most popular techniques. In this thesis, we will use MAP technique for speaker model adaptation unless otherwise is stated. In the following paragraphs, we will describe the technique assuming a diagonal covariance GMM configuration.

Given a background model and observation vectors from an enrolling speaker, $\boldsymbol{O} = [\boldsymbol{o}_1, \boldsymbol{o}_2, \boldsymbol{o}_3 \dots \boldsymbol{o}_T]$, we first compute sufficient statistics for the mixture weight, mean and variance parameters. In the above equations, $\boldsymbol{x}^2$ is used as a shorthand notation for $diag(\boldsymbol{x}\boldsymbol{x}')$.

$$n_i = \sum_{t=1}^{T} P(i|\boldsymbol{o}_t) \tag{2.36}$$

$$E_i(\boldsymbol{o}) = \frac{1}{n_i} \sum_{t=1}^{T} P(i|\boldsymbol{o}_t)\, \boldsymbol{o}_t \tag{2.37}$$

$$E_i(\boldsymbol{o}^2) = \frac{1}{n_i} \sum_{t=1}^{T} P(i|\boldsymbol{o}_t)\, \boldsymbol{o}_t^2 \tag{2.38}$$

where $P(i|\boldsymbol{o}_t)$ is the probabilistic alignment of the training vectors to i^{th} mixture component in the UBM which is computed as in Equation 2.35. Then, these statistics are used to update the old UBM parameters, $(w_i, \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)$, as follows:

$$\widehat{w}_i = \left(\alpha_i^w \frac{n_i}{T} + (1-\alpha_i^w)w_i\right)\gamma \tag{2.39}$$

$$\widehat{\boldsymbol{\mu}}_i = \alpha_i^m E_i(\boldsymbol{o}) + (1-\alpha_i^m)\boldsymbol{\mu}_i \tag{2.40}$$

$$\widehat{\boldsymbol{\sigma}}_i^2 = \alpha_i^v E_i(\boldsymbol{o}^2) + (1-\alpha_i^v)(\boldsymbol{\sigma}_i^2 + \boldsymbol{\mu}_i^2) - \widehat{\boldsymbol{\mu}}_i^2 \tag{2.41}$$

where $(\widehat{w}_i, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\sigma}}_i)$ are speaker model parameters and $(\alpha_i^w, \alpha_i^m, \alpha_i^v)$ are the adaptation coefficients for the mixture weights, means and variances respectively. The adaptation coefficients are defined data dependent as shown in Equation 2.42

$$\alpha_i^\rho = \frac{n_i}{n_i + \tau^\rho} \ where \ \rho \ \epsilon \ \{w, m, v\} \tag{2.42}$$

where τ^ρ is a fixed relevance factor and n_i is defined in Equation 2.36. Finally, the scale factor γ is set to ensure that the mixture weights satisfy the unity constraint. Although a different adaptation coefficient is used for mixture weight, mean and variance parameters in Equations 2.39-2.41, a single coefficient is preferred in most applications with a relevance factor in the range 8-20.

The data dependent adaptation coefficients control the balance between the old and new parameters. If a mixture component has a low probabilistic count, $n_i \rightarrow 0$, (i.e., the mixture is not observed in the speaker's enrollment) then the adaptation coefficient converges to zero meaning that the old UBM parameters will be emphasized in the updated model. On the other hand, for mixture components with high probabilistic count, $n_i \rightarrow \infty$, the adaptation coefficient will converge to one leading to the emphasis of the new and de-emphasis of the old parameters. This data dependent adaptation also leads to the following interpretation of the background model normalization in Equation 2.8. If a mixture component is not observed during the enrollment of the speaker, its parameters will be copied from the UBM and this mixture will produce a zero log-likelihood ratio in the authentication. As a result, the final likelihood will not be affected from the unobserved acoustic classes in the speaker's enrollment.

2.3. Decision Making

Usually, the log-likelihood ratio is compared to a pre-defined threshold for the verification decision. If the ratio is greater than the threshold, the identity claim is accepted, otherwise it is rejected.

In closed-set speaker identification, test utterance is scored against all models in the system and it is assigned to the model that yields the maximum likelihood.

$$\Lambda^* = \max_{1\leq S\leq N}\{\Lambda_S(\boldsymbol{O})\} \tag{2.43}$$

$$S^* = \operatorname*{argmax}_{1\leq S\leq N}\{\Lambda_S(\boldsymbol{O})\} \tag{2.44}$$

where N is the number of speakers in the system, Λ^* is the hypothesized speaker score, and S^* is the hypothesized speaker identity.

In open-set speaker identification, the hypothesized speaker score should also be compared to a threshold to decide whether the test sample is actually uttered by one of the registered speakers. This problem can be stated as two successive stages of identification and verification.

(i) Identification

$$\Lambda^* = \max_{1\leq S\leq N}\{\Lambda_S(\boldsymbol{O})\} \tag{2.45}$$

$$S^* = \operatorname*{argmax}_{1\leq S\leq N}\{\Lambda_S(\boldsymbol{O})\} \tag{2.46}$$

(ii) Verification

$$If\ \Lambda^* \geq \theta\ \ declare\ S^* \tag{2.47}$$

$$If\ \Lambda^{*} < \theta\ reject \tag{2.48}$$

As mentioned at the beginning, traditionally claimant speaker score is compared to a threshold for the verification decision. However, a rank-based decision making procedure might also be employed to decide on the acceptance/rejection of the claimed identity. In rank-based procedure, each authentication utterance is scored against a number of imposter models in addition to the claimant model. Then the scores are sorted in descending order. The rank of claimant model is used to make the decision. When this approach is used speaker verification task can also be viewed as closed-set speaker identification problem among the imposter and claimant models. Similar rank-based procedures are employed for speaker verification problem in previous studies (Glaeser and Bimbot, 1998; Beigi et al., 1999; Okamoto et al., 2009). In (Okamoto et al., 2009), it is claimed that the rank-based decision making outperforms comparable score normalization methods in a text-independent task. In our thesis, we will also use the rank-based decision making for verification decision and compare its performance with well-known score normalization methods in text-dependent tasks.

In order to assess the performance of different verification systems, we will use receiver operating characteristic (ROC) curves, detection error tradeoff (DET) curves and equal error rate (EER) metric. ROC curves represent the tradeoff between false rejection (FR) and false acceptance (FA) rates by changing the acceptance threshold over different operating points of the system. DET curves illustrate the same trade-off in a normal deviate scale (Martin et al., 1997). EER is another performance metric which shows the operating point where FR rate is equal to FA rate. In this thesis, the value of EER is measured from the DET (or ROC) curves and thus there might be small disturbances in the error rates due to interpolation errors.

3. CHANNEL MISMATCH COMPENSATION

In telephone-based speaker verification applications, it is well known that mismatch between background model training, enrollment and authentication channel conditions significantly degrade the verification accuracy. The performance degradation is mainly attributed to the nonlinear effects of telephone handset (Reynolds et al., 1995; Reynolds et al., 2000). Various compensation techniques are proposed in order to reduce the effects of the mismatch conditions. The techniques are mainly applied in three domains; feature domain, model domain and score domain. Feature domain techniques are applied in the front-end processing block. Model domain techniques attempt to reduce the effects of channel mismatch by enhancing speaker or background models. Score domain techniques are employed at the last stage of verification system after the log-likelihoods are calculated. In the following three sections, we will discuss the compensation techniques in their application order and review some of the well-known methods in each domain.

3.1. Feature Domain Channel Compensation

Feature domain methods are applied at the first stage of the speaker verification system. Pre-processing steps such as end-point detection and noise removal are also employed at this stage to enhance the speech quality and thus verification performance. For example, spectral subtraction methods for additive noise removal have been extensively studied in the literature (Boll, 1979; Rose et al., 1991; Ortega-Garcia and Gonzalez-Rodriguez, 1997; Hasan et al, 2004; Panda et al., 2007).

Different feature types are also compared for more robust speaker recognition. In (Reynolds, 1994), MFCCs, linear-frequency filter-bank cepstral coefficients, LPCCs and PLP cepstral coefficients are examined. In (Reynolds, 1996), effects of appending delta coefficients are investigated for mel-cepstrum features. In (Murthy et al., 1999; Orman and Arslan, 2001), new filter-bank designs are proposed in order to improve the recognition performance.

CMS (Furui, 1981), RASTA filtering (Hermansky, 1992) and frequency warping (Reynolds and Rose, 1995) are among the well-known and widely used feature domain techniques. More recently, feature warping (Pelecanos and Sridharan, 2001) and short-time Gaussianization (Xiang et al., 2002) are proposed which map the observed feature distribution to a reference distribution such as standard normal. They are used for text-independent speaker verification of cellular data in (Barras and Gauvain, 2003). Another famous feature domain method is feature mapping (FM) in which observations from different channels are mapped into a channel independent feature space (Reynolds, 2003). A generalized feature transformation is proposed in (Zhu et al., 2007) and performance is improved compared to the baseline FM. In (Mak et al., 2004), stochastic feature transformations are applied to reduce the effects of channel distortion.

Although high-level features do not actually fall under the title of feature domain channel compensation, we will discuss them in this section together with the low-level features. For text-independent applications, in (Dehak et al., 2007), continuous prosodic features are extracted using Legendre polynomials and they are modeled with JFA. In (Shriberg et al., 2005), normalized counts for each feature N-gram is modeled using SVMs. SVM and JFA based approaches are compared in (Ferrer et al., 2010) and JFA method outperformed SVM method. In both studies, prosodic contours are segmented into syllable like units using either a speech recognizer (Shriberg et al., 2005) or the valley points of energy contour (Dehak et al., 2007). Different prosodic contour segmentation alternatives are investigated in (Leung et al., 2008) and it is concluded that fixed-size contour segments without the knowledge of higher level information still provide comparable performance gain when combined with the baseline spectral systems. Higher-level feature modeling studies in text-independent applications are exhaustive and an extensive review can be found in (Shriberg, 2007).

Although high-level information sources have been extensively studied for text-independent applications, to the best of our knowledge, less effort has been made to utilize them for text-dependent tasks. In (Yegnanarayana, 2005), spectral, source and suprasegmental (pitch and duration) features are combined in a TDSU task. In the study, the baseline spectral system makes us of DTW technique. The suprasegmental features are extracted using the warping path information in the DTW algorithm. Discrete HMM state duration information is integrated into decision making in a HMM based configuration in (Charlet et al., 2000) and performance improvement is achieved with a simple fusion of the acoustic and alignment scores.

In this thesis, we investigate two feature domain compensation methods which are named as long time average spectrum (LTAS) filtering and spectral mean division (SMD). Additionally, we combine prosodic and spectral features together in order to improve the verification performance in our TDSU-D. In the following two subsections, we first present two simplifying assumptions of CMS and then show how one of these assumptions can be removed with SMD. In Section 3.1.3, we discuss FM in more detail since it is closely related to our channel compensation methods. LTAS filtering is introduced in Section 3.1.4. The details of the prosodic systems will be presented in Section 3.1.5.

3.1.1. Cepstral Mean Subtraction (CMS)

CMS is one of the earliest, simplest and most effective (and therefore one of the most widely used) feature domain methods employed to eliminate the effects of convolutive noise. While writing the filter-bank equations in Section 2.1, we did not take into account the channel effects. However, in telephone-based applications, clean speech signal, $x(n)$, is corrupted with a convolutional distortion introduced by the channel, $h(n)$. Then, the channel corrupted signal, $y(n)$, can be written as

$$y(n) = x(n) * h(n) \tag{3.1}$$

where * denotes the convolution operation. Since convolution in time domain corresponds to multiplication in frequency domain, DFT of channel corrupted speech frame is

$$Y_{t,k} = X_{t,k} H_{t,k} \tag{3.2}$$

where t is the frame index and k is the DFT index. If we define DFT of the channel for each band of filter-bank

$$H_{t,i,k} = \begin{cases} H_{t,k} \; in\ the\ band\ of\ the\ i^{th}\ filter - bank \\ 0\ otherwise \end{cases} \tag{3.3}$$

where i is the filter-bank index. As a result, a channel term could be introduced to the filter-bank energy equation in Section 2.1 as follows:

$$m_{t,i} = \sum_k w_{i,k} X_{t,k} H_{t,i,k} \tag{3.4}$$

There are two simplifying assumptions of CMS. First, we assume that the channel is constant within the band of the filter. Second, the channel does not vary over the duration of the utterance.

$$H_{t,i,k} = \; H_{t,i} = H_i \quad \forall k\colon w_{k,i} \neq 0 \tag{3.5}$$

Under these two assumptions, log filter-bank coefficients can be re-written as

$$\begin{aligned} \log m_{t,i} &= \log \sum_k w_{i,k}\, X_{t,k} H_i = \log\left(H_i \sum_k w_{i,k}\, X_{t,k} \right) \\ &= \log H_i + \log \sum_k w_{i,k}\, X_{t,k} \end{aligned} \tag{3.6}$$

Mean of log filter-bank coefficients can be calculated as in Equation 3.7.

$$\begin{aligned} \overline{m}_i &= \frac{1}{T}\sum_{t=1}^{T} \log m_{t,i} = \frac{1}{T}\sum_{t=1}^{T}\left(\log H_i + \log \sum_k w_{i,k}\, X_{t,k} \right) \\ &= \frac{1}{T}\sum_{t=1}^{T} \log H_i + \frac{1}{T}\sum_{t=1}^{T}\left(\log \sum_k w_{i,k}\, X_{t,k} \right) = \log H_i + \bar{X}_i \end{aligned} \tag{3.7}$$

where T is the total number of frames in the utterance. The second term in the above equation is denoted by $\bar{X}_i$ since it only depends on the clean speech signal. Finally, the mean in Equation 3.7 is subtracted from each frame to remove the channel term, (i.e., $\log H_i$).

$$\begin{aligned} \log m_{t,i} - \bar{m}_i &= \log H_i + \log \sum_k w_{i,k} X_{t,k} - \log H_i - \bar{X}_i \\ &= \log \sum_k w_{i,k} X_{t,k} - \bar{X}_i \end{aligned} \tag{3.8}$$

As observed in Equation 3.8, cepstral mean subtracted features are no longer dependent on the channel and thus invariant to the channel variations. However in the equation, we also observe that $\bar{X}_i$ is subtracted from each frame. $\bar{X}_i$ represents long-term average cepstrum of the clean speech in the filter-bank channel and contains some speaker-specific information (Garcia and Mammone, 1999). In fact, early studies in speaker recognition relied on similar average spectrum information as mentioned in Section 1.3. However, it is also argued in (Reynold and Rose, 1995) that this term exhibits significant intra-speaker variability over time and is also susceptible to variations due to speech effort and health. Therefore, removing it from clean speech might improve the verification performance by minimizing the intersession variability.

A simple flowchart of a speaker verification system employing CMS is shown in Figure 3.1.

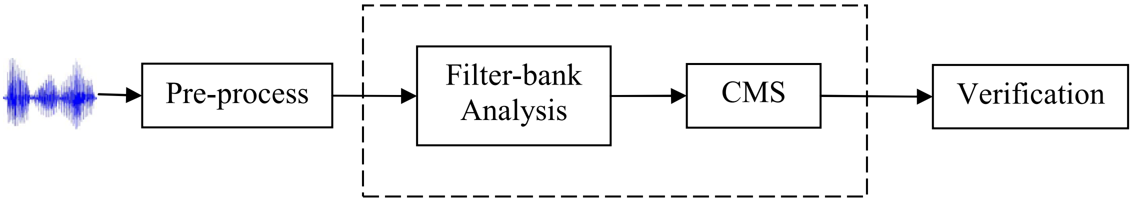

Figure 3.1. Usage of CMS in speaker verification.

The first assumption of CMS can be relaxed if we take very narrow frequency bands in which the channel becomes closer to constant. This also suggests that the normalization might be applied with the highest possible frequency resolution before the filter-bank analysis (Avendano and Hermansky, 1997; Neumeyer et al., 1994). This alternative will be discussed in the next SMD subsection.

3.1.2. Spectral Mean Division (SMD)

Although CMS is a simple and effective channel compensation technique which has been widely used in speaker verification applications, it has two simplifying assumptions as pointed out in the previous subsection. In order to eliminate the need for one of the assumptions, we present another feature domain channel compensation technique in which the normalization is performed before the filter-bank analysis in spectral domain. The method is abbreviated as SMD and can be regarded as an alternative to CMS. Moreover, it can be used together with CMS to further reduce the effects of channel distortion.

In SMD, we first compute long-term average spectrum of the channel corrupted speech signal in Equation 3.2,

$$\bar{Y}_k = \frac{1}{T}\sum_{t=1}^{T} Y_{t,k} = \frac{1}{T}\sum_{t=1}^{T} X_{t,k}H_{t,k} = H_k\frac{1}{T}\sum_{t=1}^{T} X_{t,k} = H_k\bar{X}_k \tag{3.9}$$

where we assume that the second assumption of CMS is valid, that is the channel is stationary over the duration of the utterance and $\bar{X}_k$ denotes long-term average spectrum of the clean speech at k^{th} DFT index.

Then, each frame's spectrum is divided by the average in Equation 3.9 to cancel out the constant channel term (i.e., H_k)

$$Y_k^{SMD} = \frac{Y_{t,k}}{\bar{Y}_k} = \frac{X_{t,k}H_k}{\bar{X}_kH_k} = \frac{X_{t,k}}{\bar{X}_k} \tag{3.10}$$

As can be seen from the above equation, spectral mean divided features are no longer dependent on the channel just like the cepstral mean subtracted features in Equation 3.8.

In Figure 3.2, block diagram of a speaker verification system employing SMD is shown. SMD can also be combined with CMS as shown in Figure 3.3 in order to further improve the verification performance.

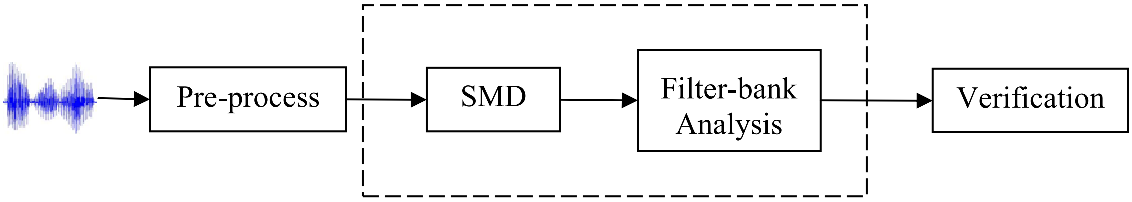

Figure 3.2. Usage of SMD in speaker verification.

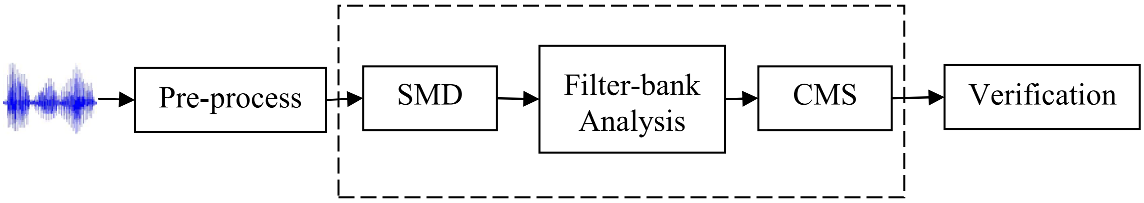

Figure 3.3. Usage of SMD+CMS in speaker verification.

3.1.3. Feature Mapping (FM)

Feature mapping (FM) was inspired by the mapping idea in speaker model synthesis (SMS) which is a model domain compensation technique and will be presented in Section 3.2. We can use Figure 3.4 to describe FM method.

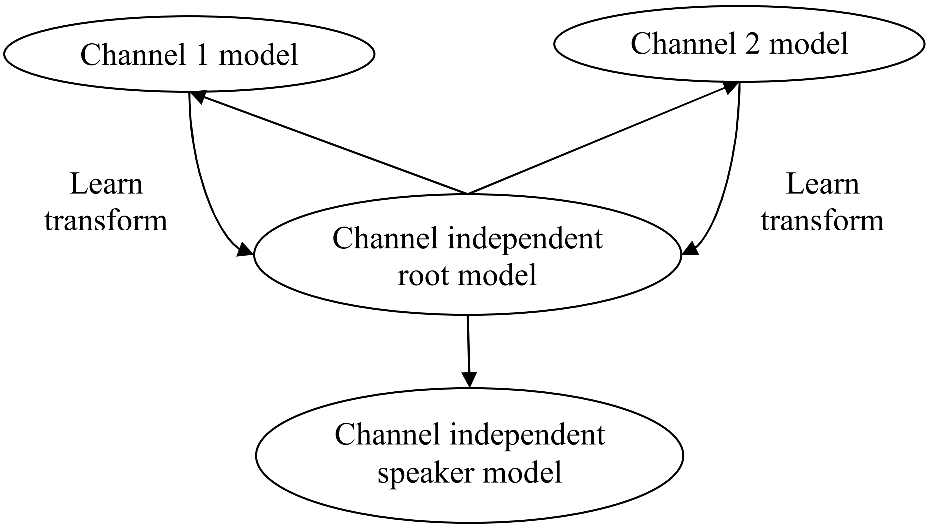
Channel 1 model
Channel 2 model
Learn transform
Learn transform
Channel independent root model
Channel independent speaker model

Figure 3.4. Feature mapping.

Since FM was first introduced for a GMM-based text-independent application, in the following derivations we will assume GMM configurations for both speaker and background models. In the offline stage of the method, first a channel independent root model is trained using speech from different speakers and channels. Then, the channel independent root model is adapted to channel dependent models using channel dependent data and MAP adaptation. When an input utterance is supplied to the system, the most likely channel type is detected for the utterance. Assume that channel 1 is detected in the channel detection step. Then top-1 decoded Gaussian in the channel 1's GMM is obtained for each feature vector ($\boldsymbol{o}_t$),

$$i = \underset{1 \leq k \leq M}{\operatorname{argmax}} \left(w_k^{CD1} N(\boldsymbol{o}_t, \boldsymbol{\mu}_k^{CD1}, \boldsymbol{\Sigma}_k^{CD1}) \right) \tag{3.11}$$

where M is the number of mixtures in the GMM, w_k is the mixture weight for k^{th} mixture and $N(:, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the normal distribution. The mapping of $\boldsymbol{o}_t = [o_{t,1}, o_{t,2}, \dots, o_{t,D}]$ to a channel independent feature, $\boldsymbol{y}_t = [y_{t,1}, y_{t,2}, \dots, y_{t,D}]$, is achieved with the following formula (where D is the dimension of observation vectors):

$$y_{t,j} = \left(o_{t,j} - \mu_{i,j}^{CD1}\right)\frac{\sigma_{i,j}^{CI}}{\sigma_{i,j}^{CD1}} + \mu_{i,j}^{CI} \qquad j = 1, 2, \dots, D \tag{3.12}$$

In enrollment, the mapped features of the enrollment utterances are used for MAP adaptation of the channel independent root GMM. In authentication, test utterance's mapped features are scored against the speaker and channel independent root models for the likelihood ratio score calculation.

As we indicate in the above procedure, channel detection should be performed to implement FM method. JFA which is the most successful channel compensation technique in recent NIST evaluations is also related to FM. However, JFA treats channel supervectors as continuous rather than discrete and eliminates the need for channel detection in a pre-processing step (Kenny et al., 2007).

3.1.4. Long Time Average Spectrum (LTAS) Filtering

The aim of LTAS filtering is to remove the mismatch due to the average spectrum of authentication and enrollment channels. This is achieved with pre-computed LTAS transformation filters. In the method, we first estimate average spectrum of each channel type by summing short time spectrum of various utterances from the corresponding channel. The utterances are taken from several different speakers to cancel out possible speaker factors. Then, LTAS transformation filter for each pair of channel is computed with a simple division. The transformation filter from an authentication to enrollment channel pair is calculated by dividing the enrollment channel's LTAS to the authentication channel's LTAS. During verification, if the authentication handset type is different from the enrollment handset, then short time spectrum of speech signal is filtered by the appropriate LTAS transformation filter. Filtered frames are overlapped and added (OLA) to reconstruct the authentication utterance.

3.1.5. Prosodic features

In this thesis, we combine prosodic and spectral features together in order to improve the verification performance in the TDSU task. We use whole-phrase sentence HMM structure as the baseline spectral system and extract duration, pitch-related (raw pitch, delta pitch and voiced-unvoiced decision based on raw pitch values) and energy features using the acoustic segmentation of the HMM states. Following five prosodic features are tested in our experiments.

(i) Pitch: For each frame, a pitch value is computed using RAPT algorithm (Talkin, 1995).

(ii) Delta-pitch: For each frame, a delta-pitch value is calculated using the raw pitch values of the preceding and following frames as shown in Equation 3.13.

$$\varDelta P(t) = \frac{P(t+1) - P(t-1)}{2} \tag{3.13}$$

(iii) Voiced-unvoiced decision: For each state, a voiced-unvoiced decision is given using the pitch values of the frames aligned to the corresponding state. We apply majority voting on the individual pitch values. A zero pitch indicates an unvoiced frame.

(iv) Energy: For each frame, a log-energy value is computed.

(v) Duration: For each state, a duration value is computed using the number of frames aligned to the corresponding state. The duration values are also normalized with the total number of frames in the utterance.

In Figure 3.5, we illustrate the alignment of feature vectors to sentence HMM states and extraction of prosodic features using the alignment information. In the figure, frame-related features (pitch, delta-pitch and energy) are represented with *F* symbol and state-related features (voiced-unvoiced decision and duration) are represented with *D* symbol. We assume that sentence HMM has 64 states.

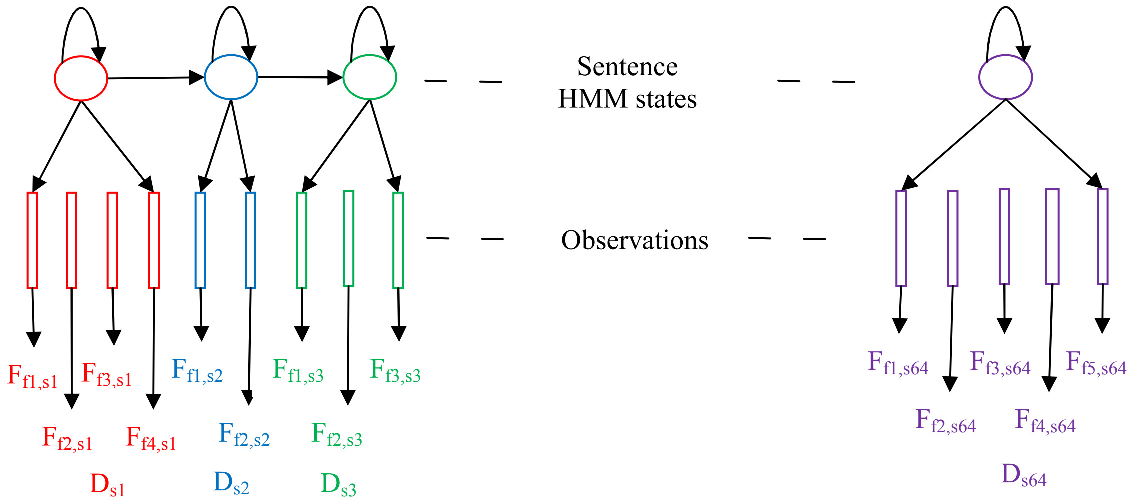

Figure 3.5. Feature extraction for prosodic systems using sentence HMM state alignment information.

In order to compare prosodic features between enrollment and authentication utterances we need a time alignment technique. We make use of state alignment information in the baseline sentence HMM method for this task. In the procedure, each enrollment utterance is aligned using the speaker's own sentence HMM adapted with the same enrollment utterances. In authentication, the claimant model is used for the alignment.

During enrollment, we compute prosodic statistics for each HMM state using prosodic values of the frames aligned to the corresponding state. Mean statistic for the frame-related features, μ_{F_j}, is calculated as in Equation 3.14.

$$\mu_{F_j} = \frac{1}{A_j}\sum_{k=1}^{E}\sum_{i=1}^{A_j^k} F_{i,j}^{k} \qquad (3.14)$$

where superscript k denotes the enrollment utterance number and E denotes the total number of enrollment utterances. A_j^k is the number of frames aligned to state j in k^{th}

enrollment utterance and A_j is the total number of frames aligned to state j during enrollment.

Mean statistic for the state-related features is computed as in Equation 3.15 where μ_{D_j} denotes the mean value for state j.

$$\mu_{D_j} = \frac{1}{E}\sum_{k=1}^{E} D_j^k \tag{3.15}$$

In authentication, the distance between enrollment and authentication utterances' prosodic features needs to be computed for the verification decision. Verification scores for the frame and state related features are calculated as shown in Equations 3.16 and 3.17, respectively.

$$\Gamma\_F = \frac{1}{A}\sum_{j=1}^{N}\sum_{i=1}^{A_j} \left|F_{i,j} - \mu_{F_j}\right| \tag{3.16}$$

$$\Gamma\_D = \frac{1}{N}\sum_{j=1}^{N} \left|D_j - \mu_{D_j}\right| \tag{3.17}$$

where N is the total number of states in the HMM, A is the total number of frames in authentication utterance and A_j is the number of frames aligned to state j. In verification score calculation we choose absolute difference after some informal trials. In (Charlet et al., 2000), the best performance is achieved with the same distance measure. In prosodic systems, we did not make use of variance statistic since we think that there is not adequate number of samples to reliably estimate the parameter. Our informal tests also verified this observation in terms of verification performance.

By deriving pitch and energy statistics from state alignment information, we can trace and compare prosodic contours using acoustically relevant speech segments.

Additionally, duration and voiced-unvoiced information of the HMM states might provide complementary information to the baseline spectral system.

3.2. Model Domain Channel Compensation

Model domain methods are applied at the second stage of the speaker verification system. They aim to reduce the effects of channel mismatch by improving the background or speaker models. However, they have a major drawback compared to feature domain techniques since they are tied to the employed classification method. Here, we can indicate that most of the methods we review in the next paragraph are studied for GMM-based text-independent applications.

The effects of training handset-dependent background models are investigated in (Heck and Weintraub, 1997) and it is observed that utilizing a handset-dependent background model which is matched to the handset type of the claimant's enrollment improves the verification performance. In (Beaufays and Weintraub, 1997), variances of speaker models are modified to compensate for channel variations. However, the use of this method requires a stereo database. Speaker model synthesis (SMS) method is proposed in (Teunen et al., 2000). In the method, a new speaker model is synthesized for the test channels which are not observed during the claimant's enrollment. As a consequence of SMS, the verification is always performed in matched conditions. Wu et al. proposed a cohort-based SMS as an enhancement to the baseline SMS (Wu et. al, 2006). In (Tadj et. al, 1999), MLLR transformation is applied for environmental compensation. In the study, speaker models are adapted to the test condition using a MLLR transformation which is computed from the test utterance. However, this MLLR adaptation procedure did not improve the performance significantly. On the other hand in (Yiu et al., 2007), pre-computed MLLR matrices are used to transform the speaker models to handset-dependent MLLR-adapted speaker models.

In this thesis, we employ MLLR transformation for compensating the effects of mismatch channel conditions. Our MLLR transformation procedure resembles (Yiu et al., 2007). However, different from (Yiu et al., 2007) we apply it in an HMM-based text-dependent framework. In the following subsection we first review SMS since it is closely related to our MLLR transformation method. Then we introduce MLLR transformation for channel compensation.

3.2.1. Speaker Model Synthesis (SMS)

Figure 3.6 can be used to describe SMS method.

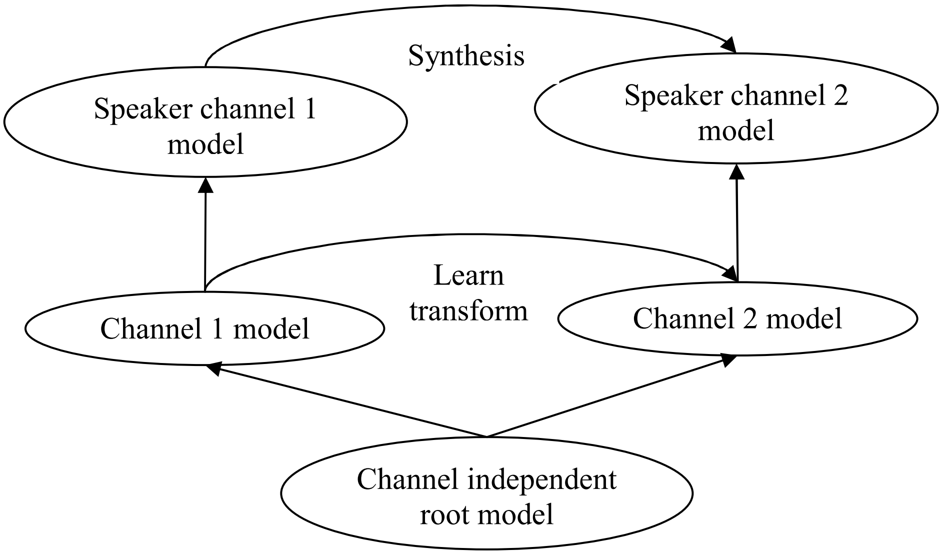

Figure 3.6. Speaker model synthesis.

Offline training stages of SMS consist of the same steps with FM. We first train a channel independent root model using speech from various different speakers and channels. Then, we adapt the root model to channel dependent background models using channel dependent data and MAP technique. Using the channel dependent background models, the following transformations for mixture weights, means and variances are obtained for each pair of channels. In Equations 3.18-3.20, we denote the transformation from channel 1 to channel 2 with $T_i^{CD1 \to CD2}$ for i^{th} mixture component. In Equation 3.20, the second subscript denotes an element of D-dimensional standard deviation vector $\boldsymbol{\sigma}_i = [\sigma_{i,1}, \sigma_{i,2}, \dots, \sigma_{i,D}]$.

$$T_i^{CD1\to CD2}(w_i) = w_i\left(\frac{w_i^{CD2}}{w_i^{CD1}}\right) \tag{3.18}$$

$$T_i^{CD1\to CD2}(\boldsymbol{\mu}_i) = \boldsymbol{\mu}_i + \left(\boldsymbol{\mu}_i^{CD2} - \boldsymbol{\mu}_i^{CD1}\right) \tag{3.19}$$

$$T_i^{CD1\to CD2}(\sigma_{i,j}) = \sigma_{i,j}\left(\frac{\sigma_{i,j}^{CD1}}{\sigma_{i,j}^{CD2}}\right) \quad j = 1, 2, \ldots, D \tag{3.20}$$

During enrollment, the most likely handset type is detected for the enrollment session and speaker model is adapted from the corresponding channel dependent background model using MAP technique. When an authentication utterance is provided to the system, its handset type is first detected. If the detected handset type does not correspond to the claimant's handset, then a new speaker model is synthesized by applying the appropriate transformation between the enrollment and authentication channels. The log likelihood ratio score is computed using the synthesized speaker model and the corresponding channel dependent background model.

Similar to FM, a channel detection system is required in SMS to identify the enrollment/authentication channel types. Eigenchannel modeling in (Kenny et al., 2007) stands in the same relation to SMS as JFA does to FM. As is the case with JFA, no channel detection is needed in eigenchannel modeling.

3.2.2. MLLR transformation

MLLR was originally developed for speaker adaptation (Leggetter and Woodland, 1995). However, it was recently applied to GMM-based text-independent speaker verification for environment adaptation (Yiu et al., 2007). MLLR is a model adaptation technique which estimates a set of linear transformations for the mean and variance parameters of a Gaussian mixture using some adaptation data. The effect of the transformation is to alter the parameters of the initial model so that the adapted model is more likely to generate the input speech. The transformation matrix, ***W***, can be decomposed into a bias vector, ***b*** and a regression matrix, ***A***

$$\boldsymbol{W} = [\boldsymbol{b}\ \boldsymbol{A}] \tag{3.21}$$

Then the adapted mean, $\widehat{\boldsymbol{\mu}}$, is computed as in Equation 3.22 where $\boldsymbol{\mu}$ is the extended mean vector with a bias offset.

$$\widehat{\boldsymbol{\mu}} = \boldsymbol{W}\,\boldsymbol{\mu} \tag{3.22}$$

In this thesis, we use MLLR transformation for channel compensation in HMM-based text-dependent framework. Our MLLR transformation procedure resembles SMS. However, roughly speaking, the SMS equations are replaced by MLLR transformations in our method. Flowchart of the procedure is shown in Figure 3.7.

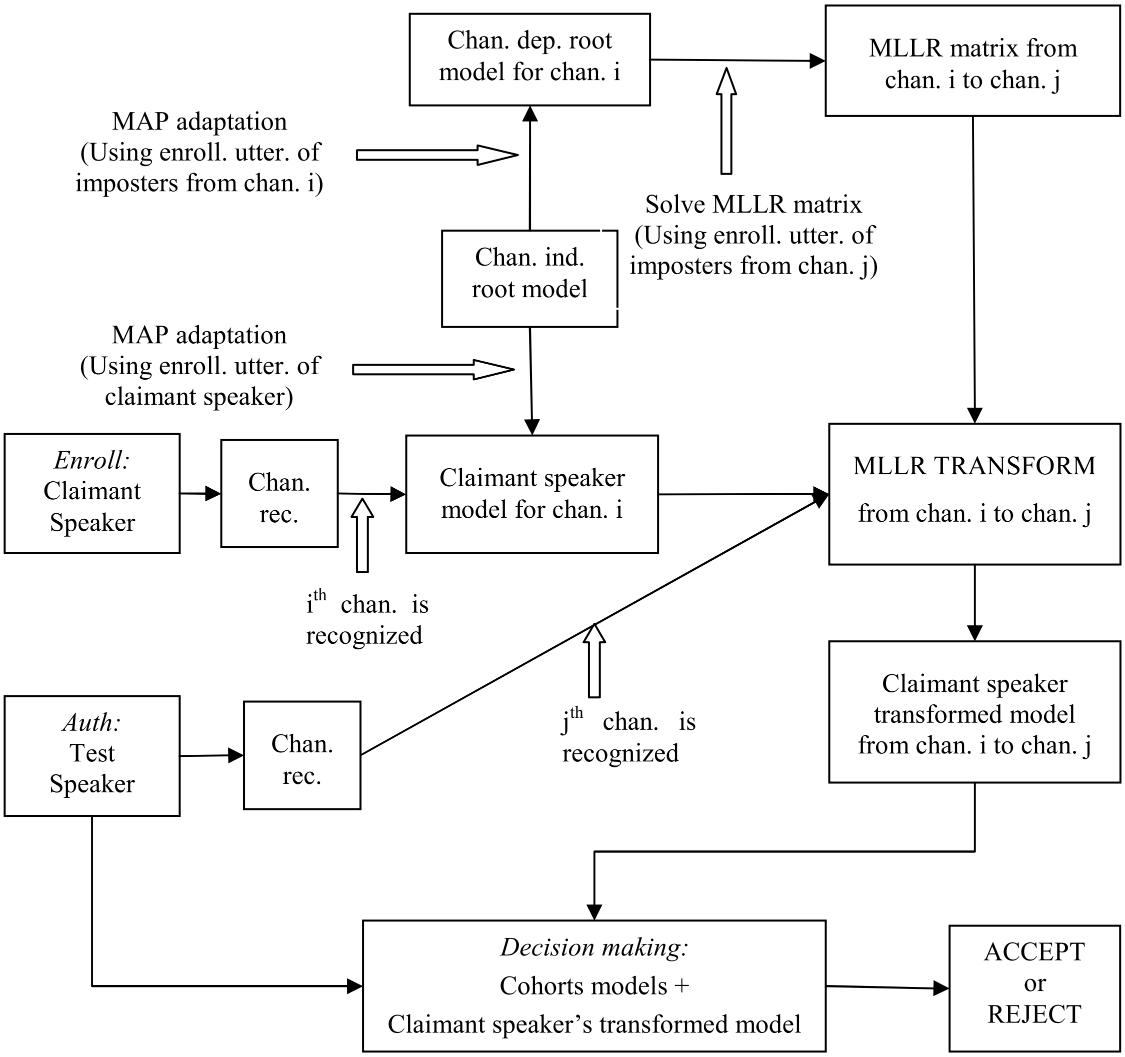
Chan. dep. root model for chan. i
MLLR matrix from chan. i to chan. j
MAP adaptation (Using enroll. utter. of imposters from chan. i)
Solve MLLR matrix (Using enroll. utter. of imposters from chan. j)
Chan. ind. root model
MAP adaptation (Using enroll. utter. of claimant speaker)
Enroll: Claimant Speaker
Chan. rec.
Claimant speaker model for chan. i
MLLR TRANSFORM from chan. i to chan. j
ith chan. is recognized
jth chan. is recognized
Claimant speaker transformed model from chan. i to chan. j
Auth: Test Speaker
Chan. rec.
Decision making: Cohorts models + Claimant speaker's transformed model
ACCEPT or REJECT

Figure 3.7. Flowchart of MLLR transformation for channel compensation.

Figure 3.7 can be briefly explained as follows. Channel and speaker independent root model is adapted to speaker independent but channel dependent root model using utterances of some imposter speakers from the specified channel. MAP adaptation is used here as in the case of SMS. Then, MLLR transformation matrices are solved for each pair of channels. The MLLR matrix from an enrollment to authentication channel pair is computed via a single iteration of Baum-Welch (BW) algorithm. The algorithm is iterated using utterances of some imposter speakers from the authentication channel and enrollment channel's root model is used as the initial model of the iteration. These two steps (obtaining channel dependent root models and solving MLLR transformation matrices) can be realized offline and are depicted at the upper side of the figure. During verification, an appropriate MLLR transformation is applied to the claimant's model if the detected handset type for the authentication utterance is different from the claimant's enrollment handset.

3.3. Score Domain Channel Compensation

Score domain methods are applied at the last stage of the speaker verification system after the speaker likelihoods are computed. Since we use claimant speaker score or rank to make the verification decision, we will divide this section into two subsections. In the first subsection, we review some of the well-known score domain methods. In the second subsection, we introduce the rank-based decision making.

3.3.1. Score-based Decision Making

Time-normalized speaker scores are compared to a threshold in score-based decision making. The speaker scores are also normalized with the background model score as in Equation 2.8. In general, the final speaker scores are further processed by some form of score normalization to compensate for non-speaker variations as shown in Equation 3.23.

$$\Lambda_{norm}(\boldsymbol{O}) = \frac{\Lambda(\boldsymbol{O}) - \mu_n}{\sigma_n} \tag{3.23}$$

where μ_n and σ_n are the estimated normalization parameters and $\Lambda_{norm}(\boldsymbol{O})$ is the normalized speaker score. In the literature, most of the proposed score normalization techniques share the normalization equation in Equation 3.23. They mainly differ in the estimation of bias (μ_n) and scale factors (σ_n). In the following paragraphs, we review two popular score domain methods, namely T-norm and Z-norm. Then, we present two possible combinations of the methods (e.g., TZ-norm and ZT-norm) and introduce a novel combination procedure referred to as C-norm. In (Gravier et al., 2000) prior knowledge of handset type was found to be useful in score normalization process similar to the use of handset-dependent background models in the background model normalization (Heck and Weintraub, 1997). Therefore, we discuss handset dependent versions of each technique together with handset independent versions.

3.3.1.1. Test Normalization (T-norm). Test normalization (T-norm) was first introduced in (Auckenthaler et al., 2000). In the normalization, each authentication utterance is scored against a set of example imposter models in parallel with the claimant model. Then, mean and standard deviation of the imposter scores are calculated. These parameters are used to perform the normalization in Equation 3.23. More recently in (Sturim and Reynolds, 2005; Ramos-Castro et al., 2007), speaker-specific selection of imposter speakers are proposed to improve the T-norm performance In handset dependent test normalization (HT-norm), the parameters are estimated using the likelihoods of cohorts who share the same channel type with the claimant's enrollment. In HT-norm, we benefit from the extra knowledge of enrollment channel type to make a better estimation.

3.3.1.2. Zero Normalization (Z-norm). Zero normalization (Z-norm) is another popular score normalization method (Li and Porter, 1988; Reynolds et al., 2000). In the method, the claimant speaker model is tested against a number of imposter utterances. Then, mean and standard deviation of the scores are calculated. These parameters are used to perform the normalization in Equation 3.23. In handset dependent zero normalization (HZ-norm), the parameters are estimated using the cohort utterances who share the same channel type with the claimant's authentication. In HZ-norm, we utilize the extra knowledge of authentication channel type to make a better estimation.

<u>3.3.1.3. Combinations.</u> T-Norm is dedicated to remove the score variability caused by the test utterances while Z-norm aims to compensate the variability due to the target models. Therefore, combination of Z-norm and T-norm might provide additional gains. In recent NIST evaluations, combination of the methods actually improved the verification performance. (Aronowitz et al., 2005; Kenny et al., 2007; Vogt et al., 2005). In this thesis, we implement two combinations named as test-dependent zero-score normalization (TZ-norm) and zero-dependent test-score normalization (ZT-norm). We also propose a novel combination procedure referred to as combined normalization (C-norm).

(i) ZT-norm and TZ-norm

We implement ZT-norm and TZ-norm as explained in (Zheng et al., 2005). In TZ-norm, Z-normalized speaker score is further normalized with the Z-normalized cohort model likelihoods as shown in Figure 3.8.

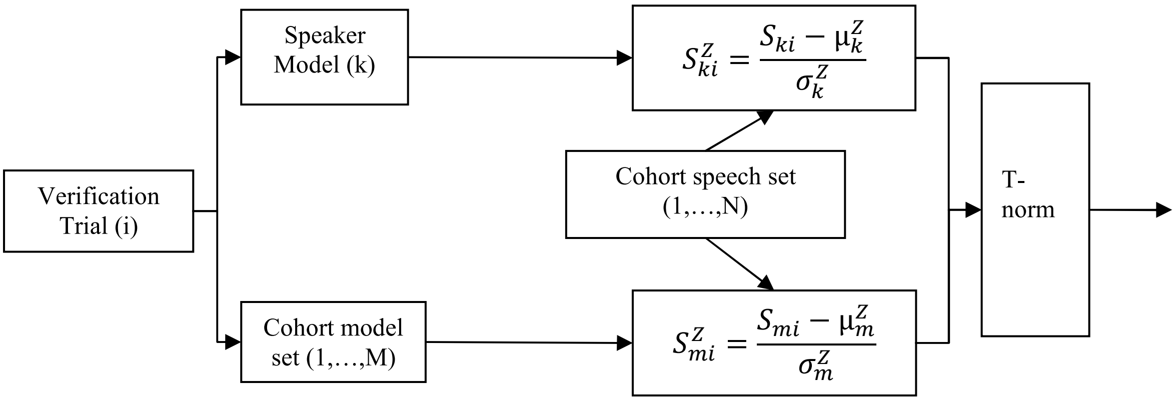
Speaker Model (k)
$S_{ki}^{Z} = \frac{S_{ki} - \mu_k^Z}{\sigma_k^Z}$
Verification Trial (i)
Cohort speech set (1,…,N)
T-norm
Cohort model set (1,…,M)
$S_{mi}^{Z} = \frac{S_{mi} - \mu_m^Z}{\sigma_m^Z}$

Figure 3.8. Schematic diagram of TZ-norm.

Similarly, in ZT-norm, T-normalized speaker score is further normalized with the T-normalized cohort utterance likelihoods as shown in Figure 3.9.

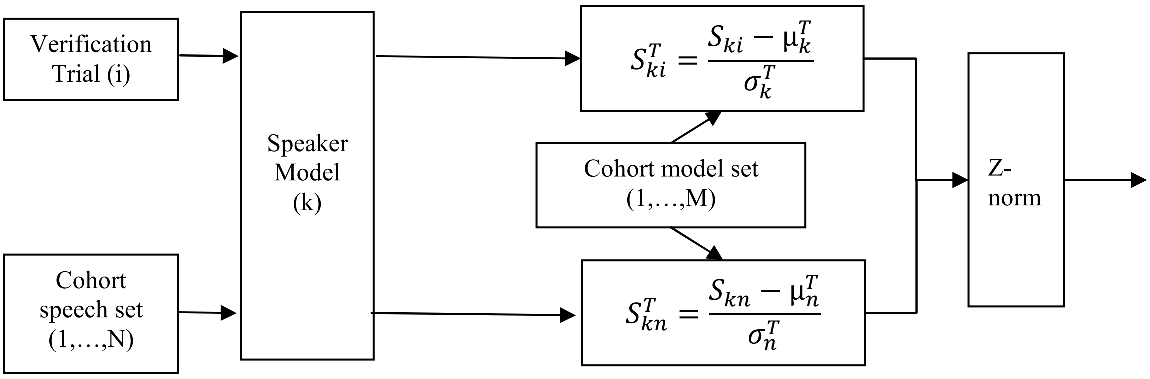
Verification Trial (i)
Cohort speech set (1,…,N)
Speaker Model (k)
$S_{ki}^{T} = \frac{S_{ki} - \mu_k^T}{\sigma_k^T}$
Cohort model set (1,…,M)
$S_{kn}^{T} = \frac{S_{kn} - \mu_n^T}{\sigma_n^T}$
Z-norm

Figure 3.9. Schematic diagram of ZT-norm.

Handset dependent versions of the normalization techniques are abbreviated as HZT-norm and HTZ-norm. In HZT-norm and HTZ-norm, we make use of prior knowledge of both enrollment and authentication channels.

(ii) C-norm

As seen from Figures 3.8 and 3.9, cohort speaker scores should also be normalized in addition to the claimant speaker score in TZ-norm and ZT-norm. This incurs extra processing and/or storage costs. Additionally, handset-independent TZ-norm and ZT-norm did not improve the performance compared to T-norm for our TDSU task. Due to these reasons, we propose the novel combination, C-norm, in which T-norm and Z-norm scores are combined directly.

In C-norm, we assume that T-norm and Z-norm scores are governed by T and Z random variables, respectively. They are assumed to be independent random variables that are normally distributed with $N(\mu_z, \sigma_z^2)$ and $N(\mu_T, \sigma_T^2)$. Then, a new random variable, C, can be defined as follows:

$$C = \frac{T + Z}{2} \sim N(\frac{\mu_Z + \mu_T}{2}, \frac{\sigma_Z^2 + \sigma_T^2}{4}) \tag{3.24}$$

We use mean and standard deviation of the new random variable in Equation 3.24 to perform the score normalization in Equation 3.23. Handset dependent version of C-norm is abbreviated as HC-norm. In HC-norm, we make use of both enrollment and authentication channel information. While estimating T-norm parameters, channel of cohort models is matched to the claimant's enrollment channel. On the other hand, in Z-norm parameter estimation, channel of cohort utterances is matched to the authentication channel.

3.3.2. Rank-based Decision Making

In this thesis, we make use of a rank-based decision making procedure in addition to the score-based decision making. We also compare the performance of rank-based method with the score domain normalization methods presented in previous subsection. In the procedure, each authentication utterance is scored against a number of cohort models in addition to the claimant model as in T-norm. Then, the scores are sorted in descending order. We use the rank of claimant model to make the verification decision. We accept the identity of the speaker if the claimant model is among the N-best matches. The threshold value, *N*, balances the trade-off between FR and FA rates. If *N* is increased, FR rate decreases at the cost of an increase in FA rate. In rank-based decision making, FA rate is strongly related to the number of cohort models $\left(\mathrm{FA} \sim \frac{\mathrm{N}}{\mathrm{numCohorts+1}}\right)$. We can also add that, as a result of the procedure the speaker verification task can be viewed as a closed-set speaker identification problem among cohort and claimant models.

For our databases, there are two alternative ways of selecting model channel of cohort speakers in rank-based method. First alternative is to use all five channel models of the cohorts. This alternative will be referred to as rank-based decision making (Rank). No channel information is required in Rank. In the second alternative, we incorporate prior knowledge of channel type to improve the verification performance and perform the ranking among the cohorts who share the same channel type with the claimant's enrollment. This choice is named as handset-dependent rank-based decision making (H-rank). Similar to HT-norm, a prior knowledge of claimant enrollment channel is required in H-rank.

4. DATABASES

In this chapter we will present our TDVT and TDSU databases but first we will make a review of other text-dependent databases in the literature. In Section 4.4, we will mention another database used for the background model training.

4.1. Text-dependent Speaker Recognition Databases

In real-life commercial text-dependent applications, enrollment and authentication phrases are chosen with respect to the requirement of the specific application. Based on the application type, all users might share a common password or a unique password can be assigned to each user. We will use the term text-dependent single utterance application (TDSU-A) for the first type and text-dependent unique password application (TDUP-A) for the second. Moreover, the system may prompt a random password from a constrained vocabulary at the time of authentication. These systems are also known as text-prompted and they are referred to as text-dependent random password application (TDRP-A) in this thesis.

Digit string is a natural choice for TDRP-A. In addition to digit passwords, personal information such as names, surnames, maiden names, place of birth, date of birth are well-suited for TDUP-A. In TDSU-A, all users repeat a single pass phrase such as “my voice is my password”, “recognize me from my voice” or “zero one two three four five six seven eight nine”. Except the TDRP-A, the other two types are more vulnerable to fraud incidents since a pre-defined password might be known in advance and played from a recorded speech. On the other hand, including the authentication utterance in the enrollment set as a whole may boost the performance of these systems.

Although NIST provides a common evaluation framework for text-independent speaker verification, there is no benchmark test for text-dependent tasks. Due to this deficiency, various researchers designed their own text-dependent databases and reported verification performances on them. Some of these databases are listed in Table 4.1. As observed in the table, text-dependent databases are collected for several different languages such as English, French, Spanish, Dutch, Chinese, Japanese, Indian, etc...

Table 4.1. Other text-dependent databases.

Database	Language	Type of utterances	Telephone or microphone	Size of database
YOHO (Higgins et al., 1991; Campbell, 1995)	English	"Combination lock" phrases (e.g., "35 - 72 – 41")	A high quality telephone handset, no telephone channel	138 speakers
CNORM + DEMO_RV1 (Huang and Rao, 1995)	English	"Combination lock" phrases (e.g., "46 - 79")	Telephone	12 speakers + 23 speakers
TUBTEL (Hardt and Fellbaum, 1997)	German	Sentences	Telephone	50 speakers
(Isobe and Takahashi, 1999)	Japanese or English	Digit strings	Telephone	25 speakers + 75 speakers
(Charlet et al., 2000)	French or English	5 short sentences	Telephone	55 target speakers 600 imposters
SESP (Bimbot et al., 2000)	Dutch	Digit strings	Telephone	46 speakers
(Lamel and Gauvain, 2000)	French	Read (digit strings, sentences) + Spontaneous speech	Telephone	100 target speakers 1000 imposters
(Rosenberg et al., 2000)	English	A single phrase	Telephone	50 male target speakers 50 male imposters
(Benzeghiba and Bourlard, 2003; Chollet et al., 1996)	French	17 words common for all speakers	Telephone	143 speakers

Table 4.1. Other text-dependent databases (continued).

(Gupta et al., 2005)	English	A phrase common to all speakers + Speaker-specific phrases	Microphone (Connected to a door control unit)	21 speakers
(Yegnanarayana et al., 2005)	Hindi (Indian)	10 fixed sentences	Microphone and telephone	30 speakers
(Nealand et al., 2005)	English	Digit strings	Telephone	354 speakers
(Hebert and Boies, 2005)	English	Fixed and random digit strings	Telephone	142 speakers
(Yuo et al., 2005)	Mandarin (Chinese)	Digit strings	Microphone	100 speakers
MVGL-AVD (Cetingul et al., 2006)	Turkish	Names + A fixed digit password	Video camera	50 speakers
(Camlikaya et al., 2007)	Turkish	“Combination lock” phrases (e.g., "35 – 45 – 66”)	Microphone	30 speakers
(Shahin, 2008)	English	8 fixed sentences	Microphone	30 speakers
(Yoma et al., 2008)	Spanish	Digit strings + First and family names	Telephone	40 speakers + 31 speakers
(Dong et al., 2008)	Mandarin (Chinese)	10 fixed passwords	Telephone	214 speakers
MSRI (Das et al., 2008)	Indian and English	Digit strings and phrases	Microphone	344 speakers
(Yamada et al., 2010)	Japanese	3 fixed sentences for enroll.-auth. + 3 fixed sentences for enroll. - 5 other fixed sentences for auth.	Microphone	10 male speakers

In the above table, YOHO has been a better known database for evaluation of text-dependent speaker recognition and it is used in various studies (Che et al., 1996; Yoma and Pegoraro, 2002; Toledano et al., 2008). It is also available from linguistic data consortium (LDC, 2011). However it lacks channel variations since the recordings are taken with a high-quality telephone handset and they are not passed through a telephone channel (Campbell, 1995). In Table 4.1, only two Turkish databases are listed. One of the databases is actually collected for audio-visual speaker recognition (Camlikaya et al., 2007) and both databases do not include multi-channel sessions. Our main objective in this thesis is to study handset-channel variability issues on text-dependent speaker verification applications. Due to the lack of an appropriate, commercially available database for the task, we collect two multi-channel speaker recognition databases which include Turkish phrases and digit strings. The first database is named as text-dependent variable text database (TDVT-D) since in its target application authentication utterances might be randomly generated from digit vocabulary. The second database consists of multiple recordings of a single Turkish phrase and it is referred to as text-dependent single utterance database (TDSU-D).

4.2. Text-dependent Variable Text Database (TDVT-D)

TDVT-D consists of the recordings of 52 speakers over five different handset-channel conditions. There are 36 male and 16 female speakers in the database. Each speaker reads nine enrolment and four authentication utterances in two separate sessions. All the recordings are taken in a noisy office environment.

Throughout the recording sessions, speakers are assisted with interactive voice response (IVR) prompts. IVR system is constructed behind a public switched telephone network (PSTN) channel with an internal office extension. We placed calls to this system from five different handset-channel conditions:

(i) Fixed wired IP phone connected to an internal office extension.

(ii) Fixed wired analog phone, PSTN.

(iii) Another fixed wired analog phone, PSTN.

(iv) Fixed cell phone, GSM network.

(v) A random cell phone (each speaker's own cell phone), GSM network.

Telephone channels in the above list can be grouped into two broad categories; landline-landline and GSM-landline. First three conditions use landline to landline telephone channels. The first condition differs from the second and third ones since it uses the office's internal telephone network. In the fourth and fifth conditions, GSM-landline connections are employed. We also use different telephone handsets for each condition. Second and third handsets are similar analog phones whereas the first handset is an IP phone. Fourth condition represents a fixed GSM handset. We use random GSM handsets in the fifth condition. Although handset and channel types vary for each recording condition, we will shortly refer to them as "channel".

Enrollment utterances consist of six digit strings and three phrases listed below. Each speaker in the database repeated the enrollment utterances from the five handset-channel conditions.

(i) Enrollment digit strings: "6-3-9-0", "1-3-5-7", "2-4-6-8", "7-6-8-1", "8-9-4-0", "3-2-5-9".

(ii) Enrollment phrases: "Ağaç ne kadar yüksek olsa da yaprakları yere düşer", "Sütten ağzı yanan yoğurdu üfleyerek yer", "Tilkinin dolaşıp geleceği yer kürkçü dükkanıdır“.

The duration of each enrollment utterance is approximately 2-3 seconds. In the enrollment set, each digit appears either two or three times and 25 of 29 letters in the Turkish language occur at least once. We prefer three Turkish common sayings as enrollment phrases in order to relieve the repetition of the utterances.

Each speaker in the database has spoken four authentication utterances from the same five handset-channel conditions.

(i) Password: A fixed password ("1-2-3-4").

(ii) Random number: A random digit string generated at test time. It consists of four digits.

(iii) Repeated phrase: Repetition of an enrollment phrase (“Sütten ağzı yanan yoğurdu üfleyerek yer”).

(iv) Place of birth: A fixed and short city name (“istanbul”).

The authentication utterances in the above list will be referred to as password, random number, repeated phrase and place of birth, respectively. Password and random number will be referred to as authentication digits when required. We intentionally choose the authentication utterances in different lengths and nature to clearly observe the effects of different phenomena on speaker verification accuracy. The durations of authentication digits and repeated phrase are approximately 2-3 seconds and the length of place of birth is less than 1 second. Authentication recordings are taken in two separate sessions in order to introduce inter-session variability to our database. First session recordings are taken right after the enrollment. In the second session, just the authentication utterances are recorded. The time interval between the sessions varies from several hours to several months depending on the availability of the speakers. Five of the 52 speakers did not attend the second session. Four of those speakers were male, one was female.

4.3. Text-dependent Single Utterance Database (TDSU-D)

TDSU-D consists of the recordings of 59 speakers over 5 different handset-channel conditions. There are 42 male and 17 female speakers in the database. Each speaker repeats a single utterance, "benim parolam ses kaydımdır”[1], in which five of the eight vowels in the Turkish language appear at least once. Since it is generally considered that voiced sounds contain more speaker-specific information compared to unvoiced sounds (Lamel and Gauvain, 2000), we think that this choice might provide additional gains. The recordings are taken in two separate sessions. In the first session, speakers repeat the utterance five times. In the second session, two repetitions are recorded. The time interval between the sessions varies from several days to several weeks depending on the availability of the speaker. Speakers are assisted with IVR prompts throughout the

[1] In English “My password is my recording”.

recordings. We record the utterances in different environments with varying background noise level. However, we should mention that most of them are taken in a noisy office environment.

IVR system is implemented behind a PSTN channel. We place calls to this system from 5 different handset-channel conditions:

(i) A fixed wired analog phone, PSTN.
(ii) The same phone in the first condition but in hands-free mode, PSTN.
(iii) Another fixed wired analog phone, PSTN.
(iv) A fixed wireless digital phone, PSTN.
(v) A fixed cell phone, GSM network.

Each condition in the above list is specifically chosen to represent distinct handset-channel conditions and background noise levels. First and third handsets are two different wired, analog phones. The second condition represents noisier environment compared to the others. However, it is also more realistic for users who are often busy at work. In the fourth condition, we use a wireless telephone handset. A fixed cellular phone is used in the fifth condition. PSTN-PSTN connections are employed for the first four conditions in the database. GSM-PSTN channel is used in the fifth one.

4.4. Background Model Training Database (BMT-D)

In some of our experiments, we use a different database for the background model training. This database is larger and phonetically more balanced compared to the speaker recognition databases and it is referred to as background model training database (BMT-D). BMT-D includes 7.8 hours of conversational and read telephone speech for landline and GSM channels. Actual call center recordings constitute approximately half of it. It is a part of telephone based Turkish speech recognition database collected internally at Sestek Inc.

5. TDVT EXPERIMENTS

In this chapter, we present our experimental results in the TDVT-D. This chapter is divided into two sections. Section 5.1 is devoted to model selection and Section 5.2 is devoted to channel mismatch compensation. We employ the rank-based procedure for decision making throughout the chapter.

5.1. Model Selection

We compare the performances of a monophone HMM based method with a baseline GMM based method in this section. CMN is used to normalize the feature vectors in the experiments. We provide the results of a closed-set speaker identification experiment in addition to speaker verification.

5.1.1. GMM Implementation

In this thesis, GMM implementation is realized with Becars toolkit (Blouet et al., 2004). 512-mixture speaker independent UBM is trained using the BMT-D. Speaker models are adapted from UBM using the speaker's enrollment utterances. MLLR-MAP technique (Blouet et al., 2004) is used to update only mean vector parameters. This adaptation procedure is chosen after informal trials. Log-likelihood ratio scores are computed for decision making.

In HMM-based method, we may train a separate model to align the long begin and end silences in the utterances. However, we do not have such an alignment property in GMM case. In order to make a fair comparison between two approaches, begin-end silences are removed from each utterance prior to feature extraction in GMM implementation. We make use of speech recognition models for the silence removal. First, phone level alignment is performed for each utterance using triphone HMMs trained with the hidden Markov model toolkit (HTK) (Young et al., 2006) for the Turkish language using approximately 50 hours of speech. Then, silence aligned segments are clipped. Feature vectors are extracted from the clipped utterances using HTK. Each vector consists of 13 MFCCs (including zeroth value) and their first order derivatives. MFCCs are computed from a sliding window of 25 ms with 10 ms frame shift.

5.1.2. HMM Implementation

We use Sphinx3 toolkit in HMM implementation (Sphinx, 2009). We first train context and speaker independent (SI) HMMs for the 29 letters in Turkish language using the utterances in the BMT-D and their corresponding transcriptions. These SI HMMs play the same role with the UBM in the GMM implementation. In addition to monophone HMMs, one silence, one noise and DTMF models are trained. All models have left-to-right topology without skip state. They have 3 emitting states and 6 mixtures.

Speaker models are adapted from SI HMM using the speaker's enrollment utterances. MAP technique is used to update only mean vector parameters. We choose the adaptation procedure after informative trials. The silence, noise and DTMF models are copied from the general model since they do not contain any speaker specific information. Therefore, each speaker model consists of speaker dependent phone models in addition to shared silence, noise and DTMF models. Forced alignment likelihoods to the known text are computed for decision making.

In HMM implementation, we try to use the same set of feature extraction parameters with the GMM case. However, we do not remove the begin-end silence segments in the pre-processing step since they are aligned with the shared silence model and do not contribute to the final speaker score considerably.

5.1.3. A Closed-Set Speaker Identification Experiment

In this subsection, we conduct a closed-set speaker identification experiment among 52 speakers in the TDVT-D. In the experiments, the threshold is set to three meaning that if correct speaker's model is among the three most likely speakers, then we assume that the speaker is identified correctly. The results are presented for all possible enrollment-authentication channel combinations in Tables 5.1-5.4. Enrollment and authentication channel conditions are indicated in the rows and columns, respectively. First five rows and columns represent the five channel conditions in the database. In each row, speaker models are adapted using enrollment utterances from the corresponding channel. Similarly, the error rates in each column are obtained using authentication utterances from the specified channel. Sixth and seventh rows represent two different multi-channel enrollment cases. In sixth row, speaker model adaptation is performed using all utterances of the speakers from five channel conditions. In seventh row, enrolment utterances from the three landline phones are used in the adaptation. Consequently, 45 utterances (9 utterances × 5 channels) in the former and 27 (9 utterances × 3 channels) utterances in the latter cases are used in the adaptation of the speaker models. These two conditions illustrate the ideal but unrealistic case in which multi-channel enrollment sessions can be carried out for each speaker. Therefore, error rates in these rows can be viewed as an upper performance bound. In the last column, average error rates for enrollment channel conditions are provided.

Since the nature of password and random number authentication utterances are similar, the individual performances for these utterances are combined in the following tables. The results in Tables 5.1-5.2 correspond to this test case. Tables 5.3-5.4 show repeated phrase error rates for HMM and GMM methods, respectively. In the tables, match condition results are shown on the diagonal entries and marked with different color. Other entries represent different mismatch conditions. The error rates for place of birth utterance are not shown with a separate table in this subsection but provided in Appendix A. In the following results, the error rates in two sessions are combined together. Separate error rates in the sessions can be found in Appendix A for all authentication utterances using HMM based method.

Table 5.1. FR% rates in HMM based method for authentication digits.

	1	2	3	4	5	Average
1	2.53	2.53	3.54	14.72	13.64	7.38
2	5.56	0.51	0.00	8.12	12.12	5.26
3	8.59	1.01	0.51	9.64	13.13	6.57
4	21.72	14.14	11.62	3.55	10.61	12.34
5	23.74	15.66	17.17	15.23	2.53	14.86
6	1.01	0.51	0.00	1.02	0.00	0.51
7	1.01	0.51	0.00	4.57	6.57	2.53

Table 5.2. FR% rates in GMM based method for authentication digits.

	1	2	3	4	5	Average
1	3.54	4.55	7.07	23.86	22.22	12.23
2	8.08	2.53	2.02	12.18	18.69	8.70
3	9.09	1.01	2.53	19.29	14.65	9.30
4	28.79	18.18	19.70	9.64	19.19	19.11
5	28.28	25.76	23.74	27.41	4.55	21.94
6	3.03	1.01	1.52	5.08	1.52	2.43
7	2.02	0.51	1.01	9.64	12.63	5.16

Table 5.3. FR% rates in HMM based method for repeated phrase.

	1	2	3	4	5	Average
1	0.00	2.02	2.02	8.08	8.08	4.04
2	0.00	0.00	0.00	3.03	3.03	1.21
3	1.01	0.00	0.00	5.05	3.03	1.82
4	4.04	3.03	2.02	2.02	3.03	2.83
5	13.13	10.10	11.11	7.07	1.01	8.48
6	0.00	0.00	0.00	1.01	0.00	0.20
7	0.00	0.00	0.00	1.01	3.03	0.81

Table 5.4. FR% rates in GMM based method for repeated phrase.

	1	2	3	4	5	Average
1	1.01	1.01	1.01	14.14	9.09	5.25
2	0.00	0.00	1.01	3.03	4.04	1.62
3	0.00	0.00	0.00	5.05	4.04	1.82
4	9.09	6.06	4.04	4.04	3.03	5.25
5	16.16	9.09	12.12	6.06	1.01	8.89
6	0.00	0.00	1.01	1.01	0.00	0.40
7	0.00	0.00	0.00	2.02	3.03	1.01

Error rates in Tables 5.1-5.4 are shown with bar graphs in Figures 5.1-5.4 to better visualize the effects of mismatch channel conditions. In the figures, horizontal axis represents the enrollment channel conditions. Different colors in the vertical axis of the graphs show the error rates in different authentication channel conditions.

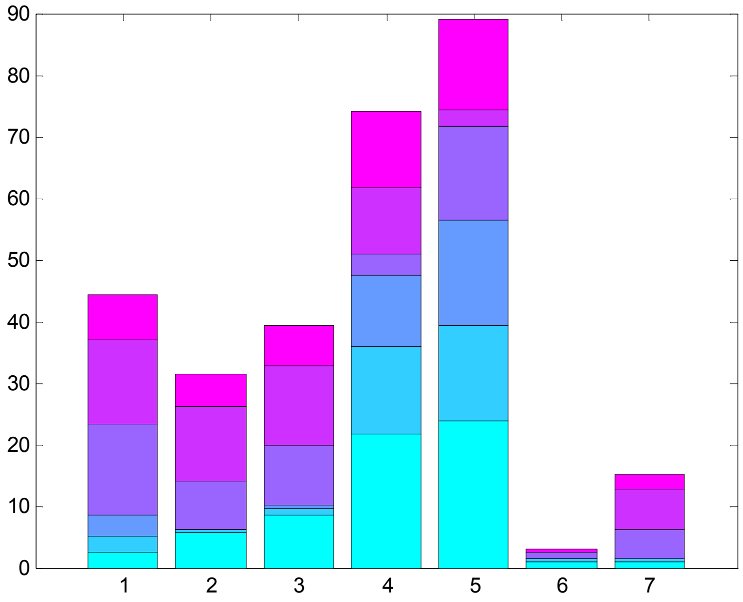

Figure 5.1. FR% rates in HMM based method for authentication digits.

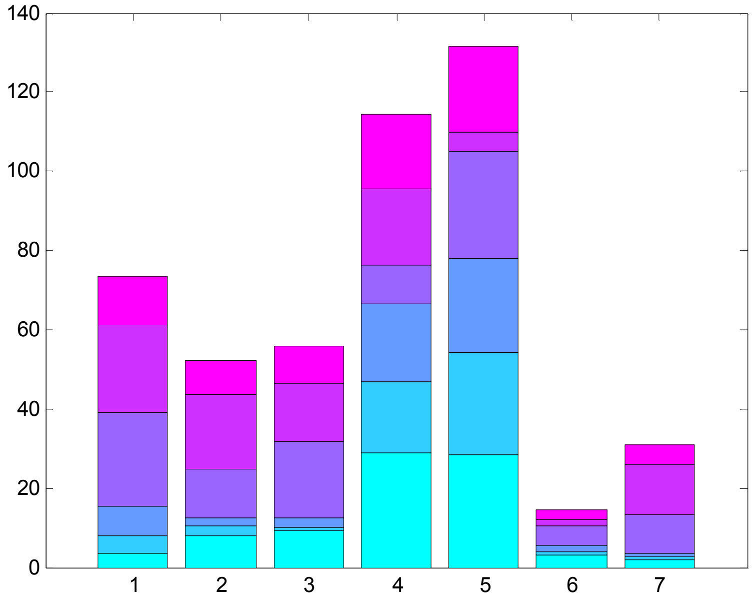

Figure 5.2. FR% rates in GMM based method for authentication digits.

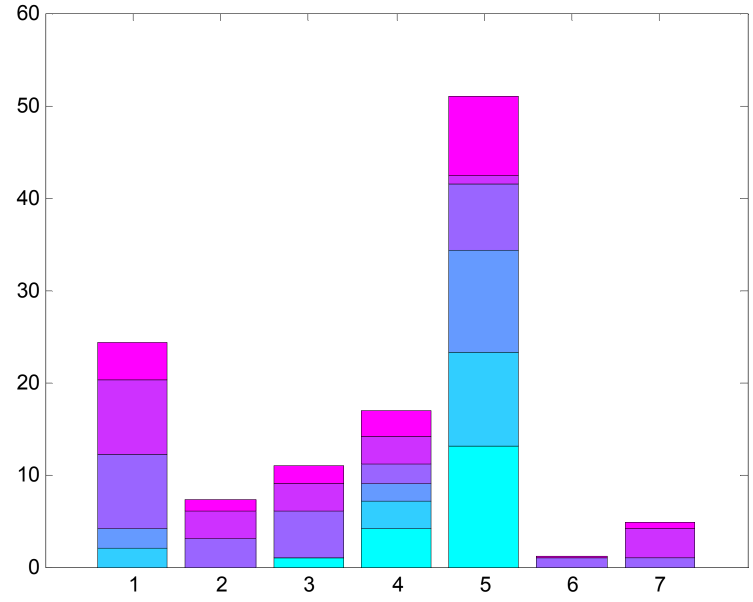

Figure 5.3. FR% rates in HMM based method for repeated phrase.

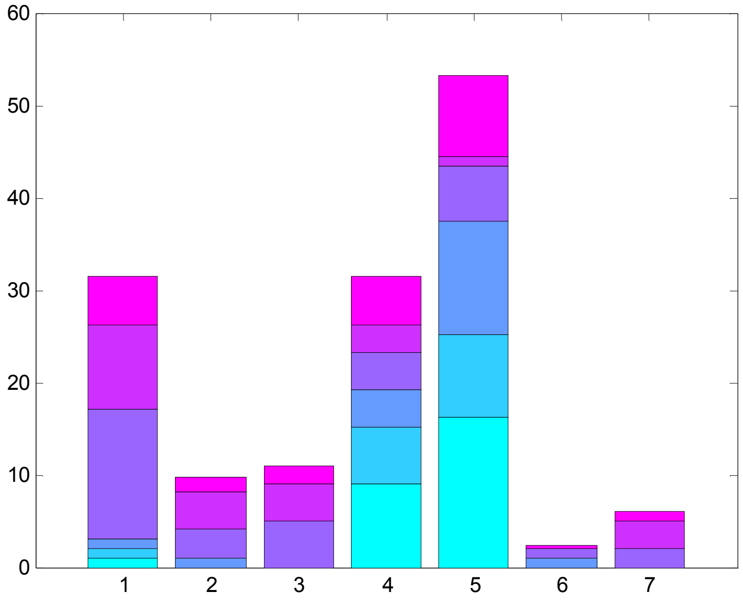

Figure 5.4. FR% rates in GMM based method for repeated phrase.

When we analyze Tables 5.1-5.4, we observe that the last two rows which correspond to multi-channel enrollment conditions show better performance compared to enrollment from single channel. The better performance in the multi-channel enrollment cases might be attributed to the larger amount of data used in the adaptation of speaker models. However, when the speaker models are adapted using enrollment speech from landline channels as in seventh row, although the performance in landline channels are very high, we cannot reach the same performance in GSM channels. This result shows that having enrollment speech from different channel conditions also has an important impact on the identification performance. On the other hand, as mentioned before multi-channel enrollment scenario is difficult to achieve in a practical setting. Therefore we will concentrate on enrollment from a single channel in the remainder of this thesis. Mismatch cases among channels 1-3 degrade the performance slightly whereas mismatch cases between landline and GSM channels show significant degradation. Mismatch between channels 2 and 3 result in minimum degradation since those two channels correspond to analog landline channels. However, using random GSM telephone handsets (channel 5) also degrades the performance to a significant extent for the mismatch case with channel 4. This can be attributed to a larger variation of microphone quality among cellular phones.

Finally, we can add that HMM based method outperforms GMM in almost all test cases in the tables.

5.1.4. Experimental Setup for Verification Experiments

Speakers in the TDVT-D are divided into two categories for speaker verification experiments. These categories are named as cohort speakers and test speakers. We perform verification trials among the test speakers. The cohort speakers are used for the normalization of target speaker scores. We prepare three different combinations of cohort and test speakers in order to increase the number of verification trials. First two sets contain 22 test speakers (14 male and 8 female). The last set contains 8 male test speakers and therefore there is no cross gender trial in this set. For all three sets, remaining 30 speakers are used as cohorts. Each speaker in the database is used only once as a test speaker and test and cohort speakers in the same set do not overlap with each other.

In verification, each authentication utterance is used as a genuine trial for its own account and as imposter trials for the other speakers' account in the test set. The trials are performed for all possible enrollment-authentication channel combinations. As a result, for the sets which contain 22 test speakers, 5 genuine (1 match + 4 mismatch condition) and 105 imposter (21 match + 84 mismatch condition) trials are carried out for each utterance. In Table 5.5, total number of genuine trials for match, mismatch and mixed conditions are given. In mixed condition, match and mismatch trials are combined together. Number of imposter trials can be calculated by multiplying the numbers in the table with 21.

Table 5.5. Total number of genuine trials.

	Match	Mismatch	Mixed
Auth. digits	989	3956	4945
Repeated Phr.	495	1980	2475
Place of Birth	495	1980	2475

5.1.5. Verification Results

In this subsection, we employ the rank-based decision making procedure among the cohort and claimant models for the verification decision. Since all model channels of the cohort speakers are used in the ranking, we do not need any prior knowledge of the enrollment/authentication channel conditions. In Figure 5.5, ROC curves for the three authentication utterances are shown. The curves are obtained for mixed condition in which match and mismatch trials are accumulated. Speaker model adaptation is performed with all enrollment utterances.

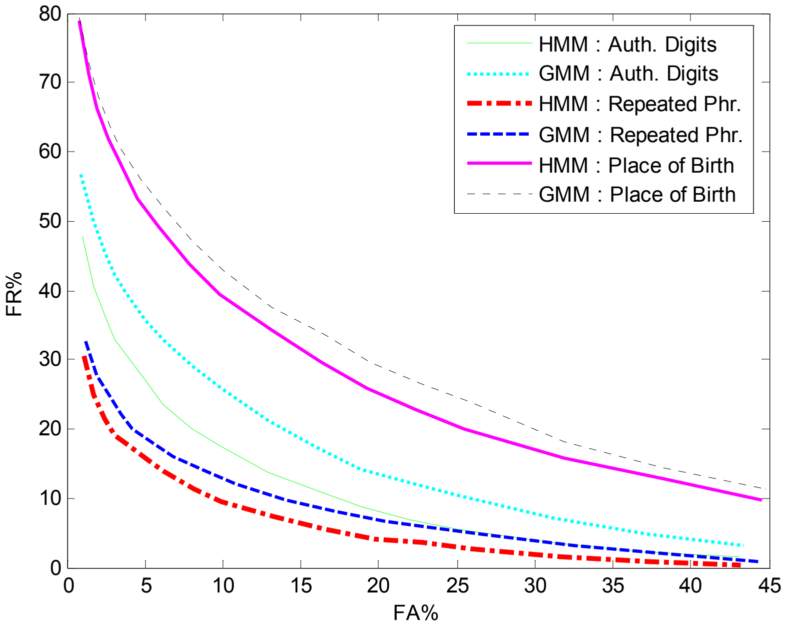

Figure 5.5. HMM and GMM methods are compared for mixed condition. Speaker model adaptation is performed with all enrollment utterances.

Several observations can be made from Figure 5.5. First, place of birth, which is relatively short utterance, results in much higher error rate compared to other test cases. This result can be attributed to the length of the authentication utterance and absence of similar context in the enrollment set. Second, HMM outperforms GMM for all authentication utterances. This is mainly because HMM captures co-articulation information better since it constrains the template matching process to the phones in enrollment and authentication sessions whereas GMM allows all training acoustic vectors to be included as viable candidates in authentication. Finally, it can be seen that inter-word temporal characteristics influence the speaker verification performance. Performance comparison in repeated phrase and authentication digits show that although they have similar length there is a significant performance gain when authentication utterance is included in the enrollment set as a whole. For authentication digits although there is intra-word match between enrollment and authentication, we do not reach the same performance level as repeated phrase.

Figure 5.5 illustrates mixed condition ROC curves. EERs for match and mismatch condition are individually shown in Table 5.6. As observed in the table, mismatch between the enrollment and authentication channel conditions significantly degrade the verification performance.

Table 5.6. EER% for match and mismatch conditions. Speaker model adaptation is performed with all enrollment utterances.

	Match		Mismatch	
	GMM	HMM	GMM	HMM
Auth. digits	5.60	3.81	16.30	13.14
Repeated Phr.	2.27	2.23	11.88	9.97
Place of Birth	14.71	14.83	24.96	23.24

In order to further investigate the effects of context match between enrollment and authentication lexicons, we prepare following three speaker model adaptation scenarios.

(i) Scenario 1: Speaker models are adapted using only repeated phrase enrollment utterance. We obtain verification results for repeated phrase authentication utterance in this scenario. This enrollment-authentication test case corresponds to single utterance speaker verification task. We will further investigate this case in Chapter 6.

(ii) Scenario 2: Speaker models are adapted using only the six digit strings in the enrollment set. In other words, the three enrollment phrases are removed from the enrollment in this scenario.

(iii) Scenario 3: Speaker models are adapted using all enrollment utterances as before.

EERs for three speaker model adaptation scenarios are shown in Table 5.7 for mixed condition. First of all, the influence of context match between enrollment-authentication lexicons is verified with the results in this table. If we compare Scenario 3 and Scenario 2, we note that the performance in repeated phrase is worsened significantly when three phrases are removed from the enrollment set. Additionally, the performance in authentication digits is better than the performance in repeated phrase in Scenario 2. Second, we can conclude that adding more data to the enrollment is not always beneficial when there is no context match with the authentication material. For example, the addition of three phrases to the enrollment set in Scenario 3 slightly degrades the performance in authentication digits when compared to Scenario 2. Similarly, although two phrases and six digit sequences are added to the enrollment set in Scenario 3 when compared to Scenario 1, the performance in repeated phrase is worse in Scenario 3. Another interesting observation is the slightly better performance of GMM over HMM for repeated phrase in Scenario 2. From this observation, we can conclude that phoneme constraint in the template matching between enrollment and authenticate sessions may degrade the performance when the phonemes in the enrollment and authentication sessions are from completely different contexts. This result also confirms the effectiveness of GMM in a text-independent task.

Table 5.7. EER% for three different speaker model adaptation scenarios for mixed condition.

	Scenario 1		Scenario 2		Scenario 3	
	GMM	HMM	GMM	HMM	GMM	HMM
Auth. Digits	-----	-----	15.95	13.01	16.64	13.37
Repeated Phr.	8.72	7.55	26.01	27.15	11.51	9.72
Place of Birth	-----	-----	31.91	32.98	24.73	22.53

5.2. Channel Mismatch Compensation

In this section, we investigate LTAS filtering, MLLR transformation and H-rank together with CMN for compensating the effects of mismatch channel conditions. The compensation methods (i.e., LTAS filtering, MLLR transformation and H-rank) are employed with/without CMN in order to assess if they provide complementary improvement with CMN. The same experimental setup in Section 5.1.4 is used in this section. HMM-based approach is preferred due to its superior performance in model selection experiments. Speaker models are adapted with all enrollment utterances. We omit place of birth authentication utterance hereafter and conduct channel compensation experiments with authentication digits and repeated phrase.

Enrollment utterances of the cohort speakers are used as development data for the compensation techniques. In the first three subsections, implementation details of the three compensation techniques are discussed. Prior knowledge of enrollment and/or authentication channel type is required in all three methods and thus following subsection is devoted to channel detection using channel GMMs. Verification results are presented in the last subsection.

5.2.1. LTAS Filtering Implementation

LTAS of each handset-channel type is estimated using enrollment utterances of the cohort speakers. In the estimation, we use a short-time Fourier analysis of 20 ms window length and 10 ms frame shift. Then, LTAS transformation filters for each pair of channels are computed as explained in Section 3.1.4. During verification, an appropriate transformation filter is applied to short time spectrum of the authentication utterance if the authentication and enrollment channel types do not match with each other. Filtered frames are overlapped and added (OLA) to reconstruct the utterance. Again, 20 ms window length and 10 ms frame shift is used in OLA. After LTAS filtering, we employ the rank-based decision making procedure among 150 cohort models (30 cohort speakers × 5 channels) and the claimant model.

5.2.2. MLLR Transformation Implementation

In MLLR transformation, the transformation matrices are computed for each pair of channels using enrollment utterances of the cohort speakers. The SI HMM in Section 5.1.2 is used as the initial speaker and channel independent root model. During verification, an appropriate MLLR transformation is applied to the claimant's enrollment model if the authentication and enrollment channel types do not correspond to each other. After the transformation, the rank-based decision making is employed among 150 cohort models and the claimant's MLLR transformed model.

5.2.3. H-rank Implementation

In H-rank, we benefit from the prior knowledge of claimant's enrollment channel type in order to perform the ranking among the cohorts who share the same channel with the claimant's enrollment. In rank-based decision making without the channel knowledge, the ranking is performed among 150 cohort models. In channel dependent version (H-rank), only 30 matched channel cohorts are used for decision making.

5.2.4. Handset-Channel Detection

We perform channel detection with Sphinx3 toolkit. We train 256-mixture channel GMMs for each handset-channel condition using the enrollment utterances of the cohort speakers. Thus, approximately 10-15 minutes of speech from 30 speakers are used in the training of each GMM. We do not make use of CMN in order to leave the convolutive telephone channel effects on the models. In recognition, a simple finite state grammar which employs equal probabilities for each channel condition is utilized.

The channel recognition performances for the test speakers' enrollment and authentication utterances are provided in Tables 5.8-5.10. In the tables, individual performances in three test sets are accumulated. Each table shows a confusion matrix, where recognized and correct channels are represented with rows and columns respectively. The recognition rates in each column are obtained using the utterances from the corresponding channel type. Diagonal entries of the matrices indicate the correct recognition rates. Off-diagonal entries represent misrecognition percentages. Table 5.8 illustrates the confusion matrix for enrollment utterances when decision on each utterance is made separately. However, during enrollment, multiple utterance decisions can be fused together to improve recognition accuracy. We apply a majority voting on individual decisions of nine enrollment utterances to decide on the enrollment channel type. The results are listed in Table 5.9. Recognition rates for the three authentication utterances (authentication digits and repeated phrase) are given in Table 5.10.

Table 5.8. Percent channel recognition rates for enrollment utterances.

	1	2	3	4	5
1	95.94	1.92	0.43	0.21	0.00
2	2.35	91.45	8.33	0.00	0.43
3	0.21	5.34	91.03	0.43	1.28
4	0.43	0.00	0.00	85.47	38.25
5	1.07	1.28	0.21	13.89	60.04

Table 5.9. Percent channel recognition rates for enrollment session. Majority voting is applied on the individual decisions of nine enrollment utterances.

	1	2	3	4	5
1	100.00	0.00	0.00	0.00	0.00
2	0.00	100.00	7.69	0.00	0.00
3	0.00	0.00	92.31	0.00	1.92
4	0.00	0.00	0.00	92.31	26.92
5	0.00	0.00	0.00	7.69	71.15

Table 5.10. Percent channel recognition rates for authentication utterances.

	1	2	3	4	5
1	94.95	4.04	0.67	1.01	0.00
2	2.36	89.23	7.07	0.00	0.00
3	0.34	5.05	91.92	0.34	1.01
4	0.67	0.34	0.00	84.12	35.02
5	1.68	1.35	0.34	14.53	63.97

When we compare Tables 5.8 and 5.10, we observe a similar performance as expected. From the tables, it can be seen that channels 4 and 5 are confused with each other more frequently. This is expected since both channels represent cellular phones. It is also observed that channels 1-3 are confused among themselves. As indicated earlier all three phones represent wired telephone handsets. We also note that misrecognition rates between channels 2 and 3 reach to 7-8%. This can be explained with the fact that both recordings are made over conventional landline channels using wired analog telephone handsets. Table 5.9 results in better performance compared to Tables 5.8 and 5.10 since it uses more data in its decision. When we analyze the results in all three tables we can say that recognition performance in fifth channel (random cell phone) is low. This channel is mostly confused with channel four. Such an error in the recognition may not be fatal since both channels correspond to GSM channels.

5.2.5. Verification Results

In order to observe the influence of employing channel detection prior to channel compensation, we perform the following ideal (channel known) and practical (channel detection) case experiments for MLLR transformation method. In ideal case, both enrollment and authentication channels are assumed to be known. Therefore no channel recognition is employed. In practical case, both channels are recognized with the channel GMMs described in the Section 5.2.4. Figure 5.6 compares ideal and practical cases for mismatch conditions in authentication digits. In Figure 5.7, the same comparison is conducted for match conditions. In both experiments, we normalize features with CMN. We did not draw the curves for other compensation techniques and repeated phrase authentication utterance since similar observations are made for these test cases in terms of channel recognition effects.

In match-ideal case in Figure 5.6, since the enrollment and authentication channels are known to be matching, no MLLR transformation is required. Therefore, the match-ideal case MLLR transformation results are identical to the baseline results. However in match-practical case, an incorrect transformation might be applied if either of the channels is misrecognized. Due to this erroneous transformation, performance might be degraded.

As seen from Figures 5.6 and 5.7, we almost obtain identical curves for ideal and practical cases. Therefore we can conclude that employing channel recognition prior to transformation does not degrade the performance compared to the ideal channel known case. This result can be attributed to the nature of channel detection errors. As observed in channel confusion matrices in Tables 5.8-5.10, most errors are due to confusion among similar channels such as landlines (channels 1, 2 and 3) and GSMs (channels 4 and 5).

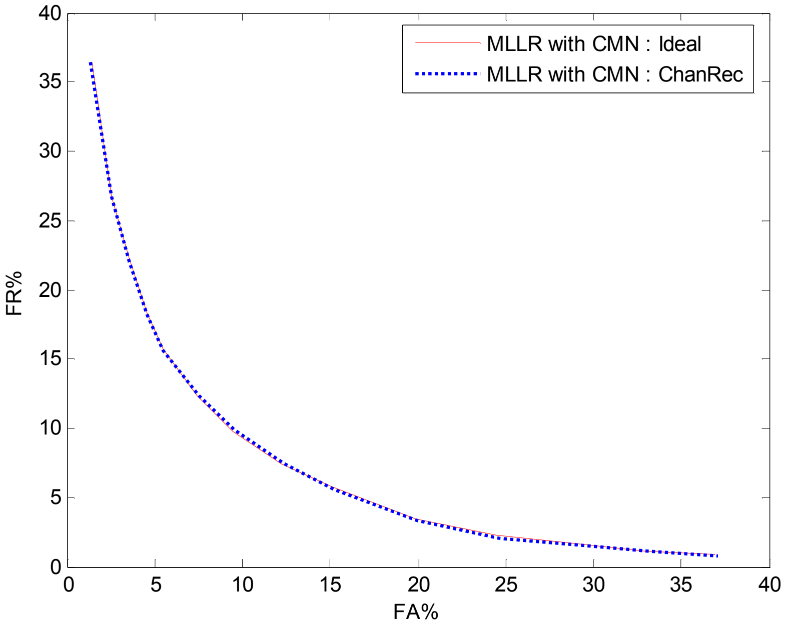

Figure 5.6. Comparison of ideal (channel known) and practical (channel recognition) cases for mismatch conditions. Verification experiment is conducted for authentication digits.

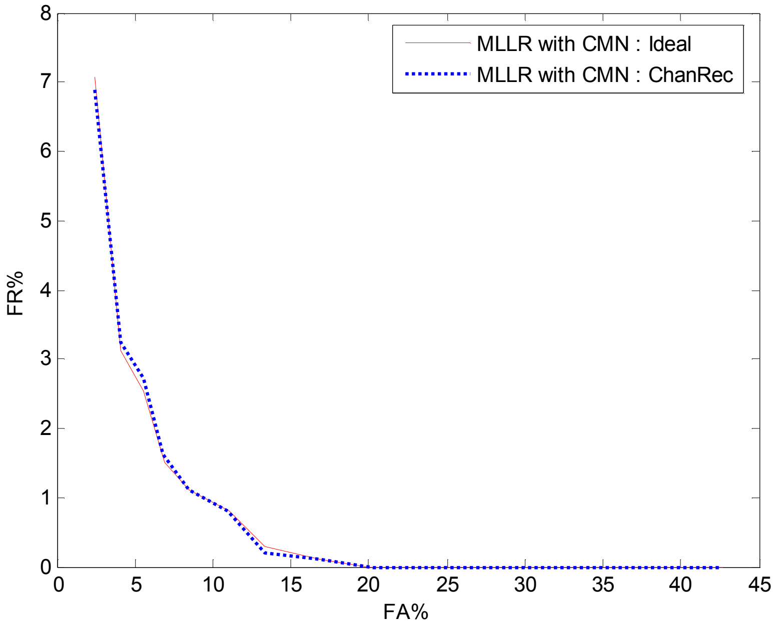

Figure 5.7. Comparison of ideal (channel known) and practical (channel recognition) cases for match conditions. Verification experiment is conducted for authentication digits.

We compare LTAS filtering, MLLR transformation and H-rank compensation techniques for authentication digits and repeated phrase utterances in Figures 5.8 and 5.9, respectively. All the curves in the figures are obtained for mixed condition. We employ channel detection prior to compensation in the experiments. The techniques are applied with and without CMN in order to assess if they provide complementary improvement with CMN. In Table 5.11, EERs in the figures are summarized.

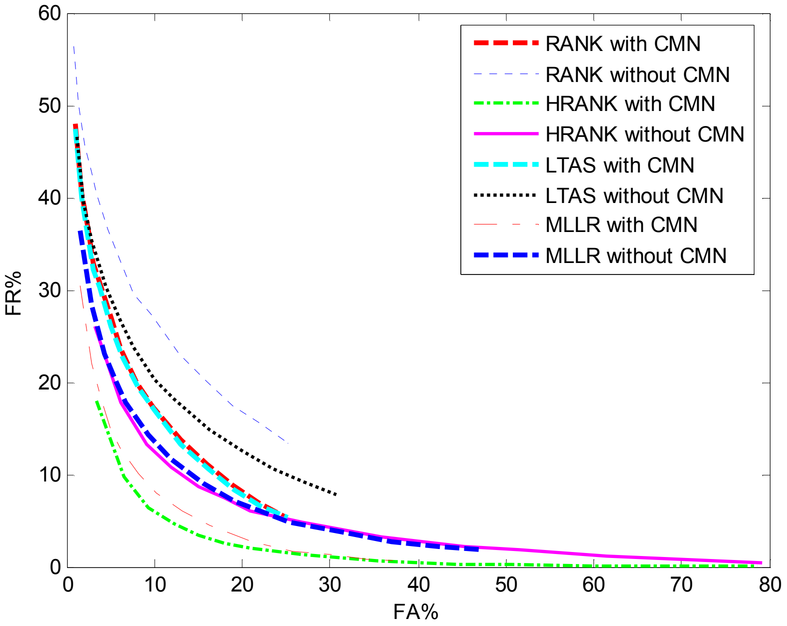

Figure 5.8. Comparison of channel compensation techniques for mixed condition. Verification experiment is conducted for authentication digits. Channel recognition is employed prior to channel compensation.

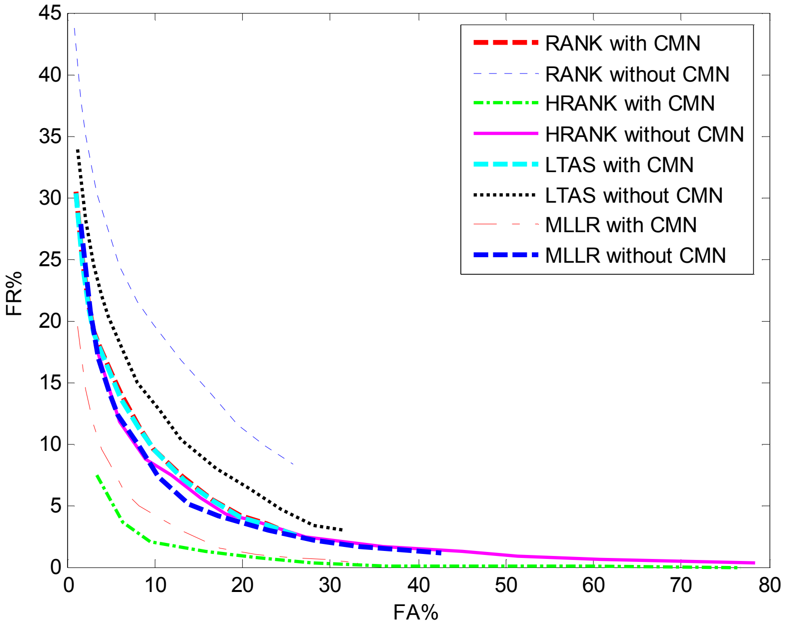

Figure 5.9. Comparison of channel compensation techniques for mixed condition. Verification experiment is conducted for repeated phrase. Channel recognition is employed prior to the compensation.

Table 5.11. EER% for mixed condition. Channel recognition is employed prior to channel compensation.

	RANK without CMN	RANK with CMN	HRANK without CMN	HRANK with CMN	LTAS without CMN	LTAS with CMN	MLLR without CMN	MLLR with CMN
Auth.digits	18.11	13.37	11.30	8.01	15.49	13.20	11.81	9.11
Rep. Phr.	15.12	9.72	8.83	5.17	11.64	9.72	8.99	6.40

We can make several observations from Figures 5.8-5.9 and Table 5.11. First, baseline experiment with CMN outperforms baseline without CMN verifying the effectiveness of CMN for channel compensation. The other three compensation techniques also perform better with CMN. LTAS filtering improves accuracy slightly when it is applied together with CMN. However, the improvement obtained with LTAS filtering is more evident when it is employed without CMN. When we compare two feature domain compensation techniques by considering the error rates in Rank with CMN and LTAS without CMN, we can conclude that CMN outperforms LTAS filtering. Among all the compensation techniques, the best performances are obtained with MLLR transformation and H-rank. This result holds for both authentication utterances and with/without CMN cases. H-rank performs significantly better than MLLR transformation in CMN case. However the performance gap between two techniques is reduced when they are used without CMN.

In the TDVT-D, the last set of experiments is performed in order to see the effectiveness of MLLR transformation and H-rank in random handset condition. For this purpose, error rates involved with fifth channel condition are accumulated. There are one match and eight mismatch cases related to the fifth channel. Channel five is the authentication channel in four of the eight mismatch cases and it is the enrollment channel in the remaining four cases. The results are presented in Table 5.12. This experiment can be regarded as a semi-open handset type experiment in which although the handsets in fifth condition are random and never observed in the offline training stages (e.g. MLLR transformation matrix estimation, channel GMM training), the other four handsets are fixed.

Table 5.12. EER% for fifth channel (random handset) condition. Channel recognition is employed prior to channel compensation.

	RANK with CMN	HRANK with CMN	MLLR with CMN
Auth. Digits	13.53	9.87	10.84
Repeated Phrase	10.20	6.35	7.79

Fifth channel condition simulates a more realistic setting where the callers speak from their own cellular phones. When we analyze Table 5.12 we observe that both compensation techniques improve the performance for the semi-open handset type experiment in this channel. This result can be attributed to the fixed GSM handset in the fourth condition and other random GSM handsets in the fifth condition which may show similar characteristics to the random GSM handset used in the test. The result also leads us to think that if we introduce more channel conditions to the database, we may obtain better performance for a random handset-channel combination in practice.

6. TDSU EXPERIMENTS

In this chapter, we present our experimental results in the TDSU-D. This chapter is divided into model selection and channel mismatch compensation sections similar to Chapter 5.

6.1. Model Selection

We compare the performances of a GMM and two HMM based techniques to find the most appropriate classification method for the TDSU task. In the first HMM-based approach, 3-state context independent HMMs are trained for each phoneme in the utterance. In the second approach, a single whole-phrase sentence HMM is matched to the fixed utterance to better capture co-articulation information. All three techniques share the same front-end processing module. In the first subsection, we present front-end processing block. The next three subsections are devoted to the three classification methods. In the last two subsections, experimental setup and verification results are presented.

6.1.1. Front-end Processing

Front-end processing block includes silence removal and feature extraction stages. Especially in sentence-HMM method, varying length begin-end silences might result in alignment problems and lead to significant degradation in verification accuracy. In order to avoid this problem, the silence sections are removed from each utterance prior to training and testing. We make use of the speech recognition models in Chapter 5 for the silence removal. First, phone level alignment is performed for each utterance using the triphone HMMs. Then, silence aligned segments at the beginning and end of the utterances are clipped. The common feature vectors for the three classification methods are extracted from the clipped utterances. Each feature vector consists of 13 MFCCs (including zeroth value) and their first order derivatives. MFCCs are computed for 25 ms window length and 10 ms frame shift. They are normalized using CMS. HTK toolkit is used in the feature extraction.

6.1.2. GMM Implementation

GMM implementation is realized with Becars toolkit. In order to observe the effects of different background modeling schemes, we train two separate speaker independent (SI) UBMs for two separate experiments. In the first experiment, various recordings of the fixed utterance are used to train the UBM. The training recordings are taken from several speakers and channels. This case will be referred to as GMM-UBM(Sen) in the following subsections. In the second experiment, BMT-D is used in the UBM training. We will refer this case as GMM-UBM(Gen). Both UBMs have 256 mixtures. In both experiments, speaker models are adapted from UBM using MLLR_MAP technique. Only mean vector parameters are updated in the speaker model adaptation. Log-likelihood ratio scores are used for decision making.

6.1.3. Monophone HMM Implementation

In this chapter, both HMM based methods are realized with HTK toolkit. In monophone HMM method, context and speaker independent HMMs (SI HMMs) are trained for each phoneme in the fixed utterance. SI HMMs play the same role with the UBM in the GMM implementation. We use the application specific database in the GMM-UBM(Sen) experiment for the SI HMM training. Each model has left-to-right HMM topology with 3 emitting states and 4 mixtures. Speaker dependent models are adapted from SI HMMs using the MAP technique. Only mean vector parameters are updated in the speaker model adaptation. Forced alignment likelihoods to the fixed text are used in decision-making.

6.1.4. Sentence HMM Implementation

In sentence HMM method, a single whole-phrase SI HMM is constructed for the fixed utterance. We again use the application specific database in GMM-UBM(Sen) experiment for the SI HMM training. The number of states in the model is chosen to be proportional with the number of phonemes in the utterance. In addition, the model size is kept consistent with the other two approaches. Therefore, we prefer left-to-right HMM topology with 64 emitting states and 4 mixtures. In sentence HMM method, we use MAP technique for speaker model adaptation and forced alignment likelihoods for decision making as in the case of monophone HMM method. The major difference of two HMM based methods is in the modeling of multiple occurrences of the same phonemes. In monophone HMM method, these phonemes are modeled with a single context independent HMM. However, there is no such constraint in sentence HMM technique.

6.1.5. Experimental Setup

Speakers in the TDSU-D are divided into three categories for speaker verification experiments. These categories are named as background speakers, cohort speakers and test speakers. Background speakers are set aside for speaker independent model training and consist of ten speakers. Forty speakers are used as cohorts to perform score normalization on the final likelihoods. Verification experiments are conducted with the remaining test speakers' authentication utterances. We prepare six different combinations of cohort and test speakers in order to increase the number of verification trials. Five of the sets contain nine test speakers and the last set contains the remaining four speakers. Each speaker in the database is used only once as a test speaker and test and cohort speakers in the same set do not overlap with each other. For all test sets, the same background speakers are used.

In verification, each authentication utterance is used as a genuine trial for its own account and as imposter trials for the other speakers' account in the test set. The trials are performed for all possible enrollment-authentication channel combinations. As a result, for the sets which contain nine test speakers, 5 genuine (1 match + 4 mismatch condition) and 40 imposter (8 match + 32 mismatch condition) trials are carried out for each utterance. We summarize total number of verification trials in Table 6.1. Unless otherwise stated, results in this section are provided for mixed condition in which match and mismatch trials are combined together.

Table 6.1. Total number of genuine and imposter trials.

	Genuine	Imposter
Match	951	7228
Mismatch	3804	28912
Mixed (Match + Mismatch)	4755	36140

6.1.6. Verification Results

For the following set of results, we use all utterances of background speakers from five channel conditions to train the speaker independent models in GMM-UBM(Sen) and HMM based methods. Only in GMM-UBM(Gen) experiment, BMT-D is used for UBM training as mentioned before. Speaker models are adapted from speaker independent models using three utterances from the first session. The same adaptation procedure is employed for cohort and claimant models. Verification experiments are conducted with the remaining recordings of the test speakers.

In Table 6.2, EERs of the classification methods are provided for no score normalization (speaker scores are only normalized with SI model score) and T-norm cases. We do not assume any prior knowledge of the channel types in this subsection. Therefore, all five channel models of the cohorts are used to estimate the T-norm parameters. DET curves without score normalization and with T-norm are shown in Figure 6.1 and Figure 6.2, respectively.

Table 6.2. EER% for classification methods.

	Sentence HMM	Monophone HMM	GMM-UBM (Sen)	GMM-UBM (Gen)
No-norm	2.19	2.54	3.83	5.22
T-norm	2.02	2.57	3.72	3.89

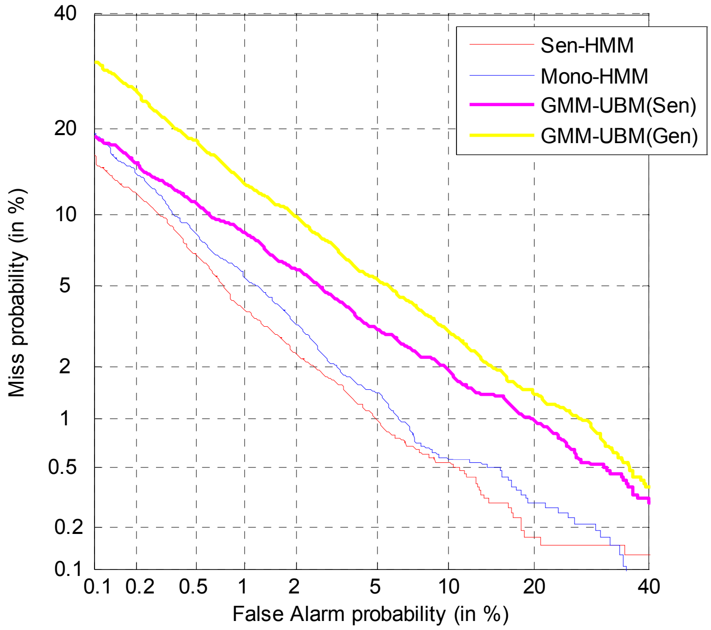

Figure 6.1. DET curves for classification methods without score normalization.

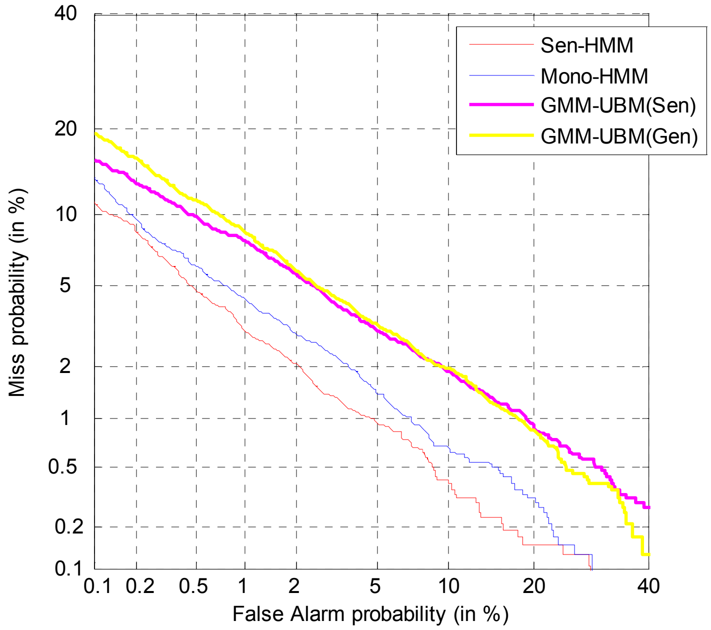

Figure 6.2. DET curves for classification methods with T-norm.

We can make several observations from Table 6.2 and Figures 6.1-6.2. First, the results in GMM-UBM(Sen) and GMM-UBM(Gen) experiments indicate that there is a performance gain in training UBM from an application specific database. Second, further score normalization is required when a general database is used to train the UBM. As observed in the last column of Table 6.2, T-norm improves the verification performance significantly in GMM-UBM(Gen) case when compared to the other three modeling configurations. Third, both HMM based methods outperform GMM. Among the four approaches, sentence HMM yields the best performance. This result can be attributed to the better modeling of the co-articulation information in the sentence HMM method.

The error rates in Table 6.2 are given for mixed condition in which match and mismatch trials are accumulated. Separate error rates in match and mismatch conditions are provided in Table 6.3 for sentence HMM method. As observed in the table, mismatch conditions degrade the verification accuracy significantly.

Table 6.3. EER% for match and mismatch conditions.

	Match	Mismatch
Sentence HMM (No-norm)	0.63	2.00
Sentence HMM (T-norm)	0.32	1.71

In order to observe the effects of enrollment data amount on verification performance, we setup another experiment in which speaker models are adapted using one, two and three enrollment utterances. The same set of authentication utterances are used in each enrollment scenario. We employ sentence HMM method with T-norm in the experiment. Figure 6.3 shows DET curves for the three enrollment scenarios. EERs in Figure 6.3 are provided in Table 6.4. As seen from the figure and table, verification performance increases when we add more utterances to the enrollment set. However, the rate of improvement decreases as the number of enrolment utterances increase.

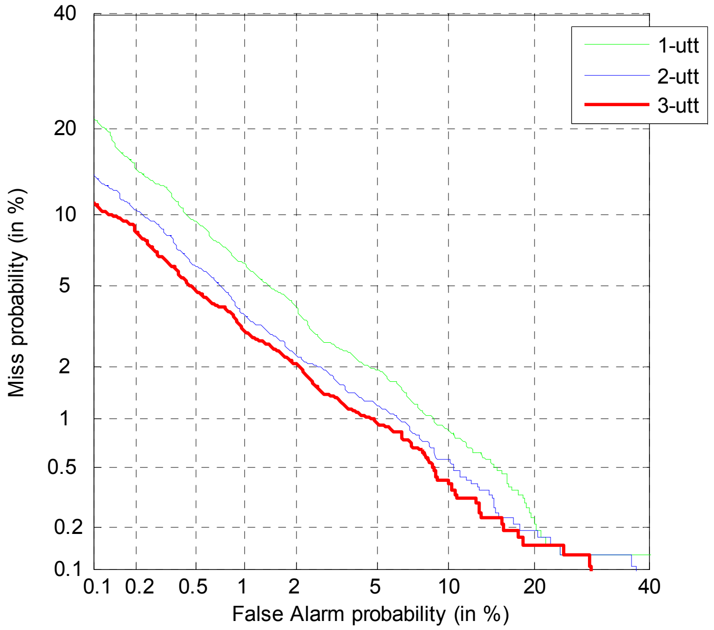

Figure 6.3. Effects of enrollment data amount. One, two and three enrollment utterances are used in speaker model adaptation.

Table 6.4. Effects of enrollment data amount on verification performance.

	1-utt	2-utt	3-utt
Sen-HMM(T-norm)	2.69	2.15	2.02

6.2. Channel Mismatch Compensation

As observed in the model selection experiments, there is a significant performance gap between match and mismatch condition results. Especially severe mismatch cases such as landline analog phone in hands-free mode in channel 2 and cellular phone in channel 5 result in high error rates. Additionally, it becomes more difficult to set a global threshold when match and mismatch trials are accumulated. This can be seen from Table 6.2 and Table 6.3 where mixed condition error rates are worse than both match and mismatch condition error rates. Therefore, we devote this section to channel mismatch compensation in order to mitigate the effects of channel mismatch. Based on the results in model selection experiments, we decided to use sentence HMM method. We adapt the speaker models with three enrollment utterances. The same experimental setup in Section 6.1.5 is used in this section.

This section is organized as follows. In the first subsection, we compare various score normalization methods for TDSU task. We present the performance of the feature domain SMD method in the next subsection and compare it with the standard CMS. The last subsection is devoted to prosodic systems and their combination with the baseline spectral system.

6.2.1. Score Normalization

We investigate the score domain normalization techniques in Section 3.3 in order to compensate for mismatched channel conditions. These normalization methods are Rank, T-norm, Z-norm, TZ-norm, ZT-norm, C-norm and their handset dependent versions. We briefly discuss implementation details of each normalization technique in the following five subsections. In handset dependent normalization, prior knowledge of enrollment/authentication channel types is required. Generally, channels are identified with channel GMMs trained for each condition (Reynolds et al., 2000; Mak et al., 2002; Mak et al., 2004). In this thesis, we present a cohort-based channel detection procedure in addition to the GMM based method. Sixth subsection is devoted to channel detection methodology. In the last two subsections, we will provide the experimental results without/with channel detection.

6.2.1.1. T-norm Implementation. In T-norm, all channel models of the cohorts are used in the estimation. Therefore, T-norm is performed with 200 models (40 cohort speakers × 5 channels). In HT-norm, the parameters are estimated using the likelihoods of cohorts who share the same channel type with the claimant's enrollment. As a result, HT-norm is performed with 40 matched channel cohort models.

6.2.1.2. Z-norm Implementation. In Z-norm, all enrollment utterances of the cohorts from five channel conditions are used to estimate the normalization parameters. Therefore, Z-norm is performed with 600 utterances (40 cohort speakers × 3 enrollment utterances × 5 channels). In HZ-norm, the parameters are estimated using the cohort utterances which are recorded from the same channel condition with the claimant's authentication. Therefore, HZ-norm is performed with 120 matched channel cohort utterances.

6.2.1.3. Rank Implementation. In Rank, all channel models of the cohorts are used in the ranking. Therefore, the ranking is performed with 200 cohort models. In H-rank, only the cohort models which share the same channel type with the claimant's enrollment are used in the ranking. As a result, HT-norm is performed with 40 matched channel cohorts.

6.2.1.4. ZT-norm and TZ-norm Implementation. In ZT-norm and TZ-norm, we use the cohort speakers' enrollment utterances for Z-norm and enrollment models for T-norm. Therefore, while Z-normalizing a cohort model likelihood in TZ-norm, the enrollment utterances of the specified cohort are not used. Similarly in ZT-norm, the enrollment model of the specified cohort is not used in T-normalization of cohort utterance likelihoods.

6.2.1.5. C-norm Implementation. The cohort set in T-norm and Z-norm are used to estimate C-norm and HC-norm parameters.

6.2.1.6. Channel Detection Methodology. In a practical speaker verification application, it is not realistic to assume that enrollment and/or authentication channels of the speakers are known. When the channel types are not known, they can be identified with a channel detection system. For this purpose, we employ two separate techniques which are named as GMM-based and cohort-based channel detection. In the following paragraphs, we first mention these two methods. Then, we present the score level fusion of the methods.

(i) GMM-based Channel Detection

In classical GMM-based method, 128-mixture GMMs are trained for each handset-channel condition using the recordings of the background speakers. In the training of each channel GMM, approximately 70 repetitions (10 speakers × 7 utterances) of the fixed utterance are used.

(ii) Cohort-based Channel Detection

In our T-norm experiments, we observed that cohort model likelihoods are boosted when the cohort speaker's channel is matched to the authentication channel type. Inspired from this observation, we propose a channel detection procedure in which cohort likelihoods from the same channel condition are averaged to estimate each channel score. Then, the channel decision is made based on those averages. To identify the enrollment channel type, each enrollment utterance is also scored against the same cohort models.

(iii) Fusion of Cohort and GMM based Methods

We also perform fusion of cohort and GMM based methods in order to improve channel detection accuracy. Combined likelihood for each channel type is simply calculated as follows:

$$L_{fuse}^{i} = (1-\alpha) * L_{Cohort}^{i} + \alpha * L_{GMM}^{i} \tag{6.1}$$

where L_{fuse}^{i}, L_{Cohort}^{i} and L_{GMM}^{i} are the fusion, cohort and GMM likelihoods for channel i, respectively and α is a scale factor. After some informal trials, we set α to 0.4. However, we observe that the detection accuracy is not very sensitive to the choice of the scale factor. We obtain close performances for values between 0.4 and 0.6. In Figure 6.4, block diagram of the fusion process is shown;

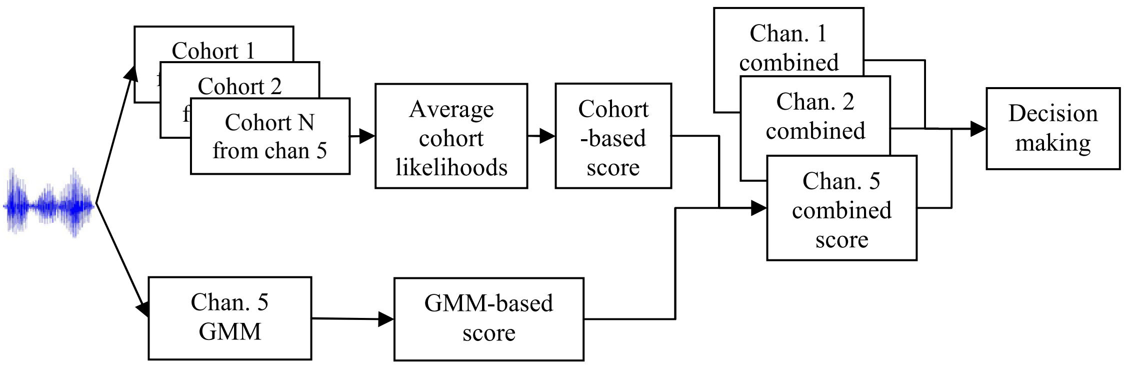

Figure 6.4. Fusion of cohort and GMM based channel detection procedures.

6.2.1.7. Verification Results without Channel Detection. In Table 6.5, EERs for the score normalization techniques are provided. In this subsection, we assume that enrollment and/or authentication channel types of the speakers are known. The effects of employing channel detection to identify the channel types will be discussed in the following subsection. In the baseline experiment in the table, speaker scores are only normalized with SI model likelihood.

Table 6.5. Comparison of score normalization techniques.

	Base	Rank	T-norm	Z-norm	TZ-norm	ZT-norm	C-norm
Handset independent	2.19	2.36	2.02	2.25	2.23	2.23	2.02
Handset Dependent	2.19	~1.43	0.86	1.05	1.32	0.78	0.72

The first observation from the table is that rank-based decision making performs worse than the other score normalization techniques. Additionally, we cannot measure the exact EER in H-rank as indicated in the table. FA rate in H-rank starts from 1.73% when acceptance threshold is set to one. This rate might be expected since forty cohorts are used in the ranking. The worse performance of rank-based decision making might also be attributed to small number of cohort speakers. Using more cohorts in the procedure might enhance the accuracy. Based on these experimental results, we omit the rank-based decision making in the remainder of this section. In Figure 6.5, DET curves for handset independent score normalization techniques are presented. In Figure 6.6, handset dependent normalization techniques are compared.

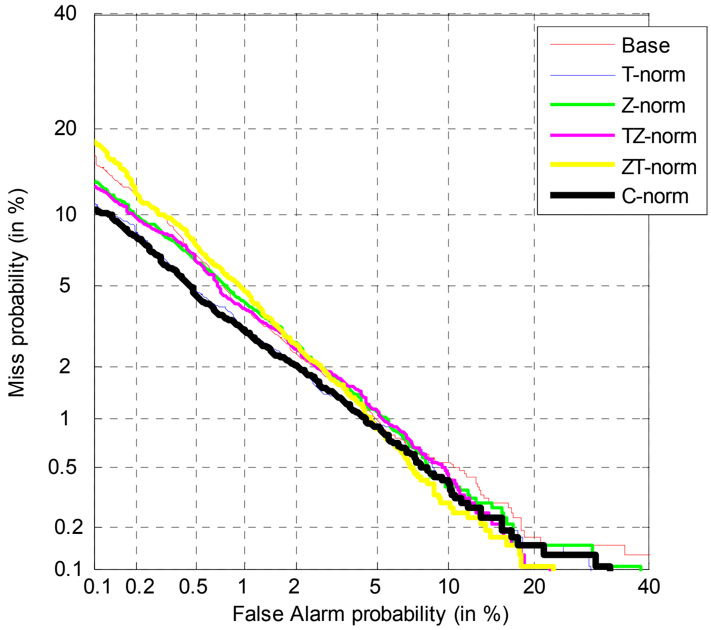

Figure 6.5. Comparison of handset independent score normalization techniques.

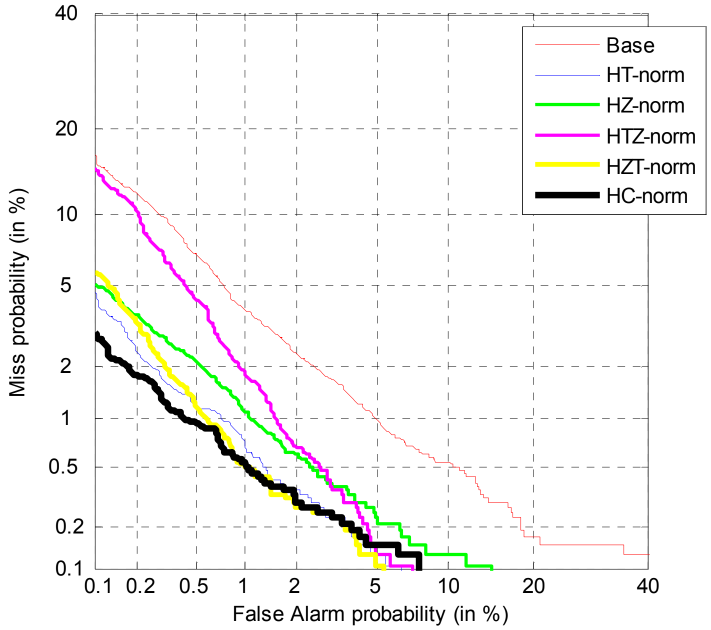

Figure 6.6. Comparison of handset dependent score normalization techniques.

We can make several observations from Figures 6.5-6.6 and Table 6.5. First, providing handset-channel information to the score normalization process significantly improves the verification performance. When channel information is used in score normalization, approximately 64-65% relative improvement in EER is achieved in C-norm and TZ-norm. Relative improvement in ZT-norm is 40%. It is approximately 53-57% in Z-norm and T-norm. Second, T-norm outperforms Z-norm in both handset dependent and handset independent cases. Among the combined techniques, the best performances are achieved with C-norm. For handset independent methods, T-norm and C-norm give the lowest error rates. HC-norm yields the best performance among the handset dependent methods. HZT-norm also performs as well as HC-norm for this case. However, as observed in Figure 6.6, HC-norm significantly outperforms HZT-norm especially in the low false alarm region.

6.2.1.8. Influence of Channel Detection Errors. Percent channel recognition rates of the three channel detection procedures are provided in Table 6.6. In the first two lines of the table, recognition rates for authentication and enrollment utterances are given. The rates are obtained for 951 authentication and 735 enrollment utterances, respectively. In identifying the enrollment session's channel type, multiple utterance decisions can be fused together to achieve a better recognition accuracy. We apply a majority voting on individual decisions of three enrollment utterances. In case of equality in majority voting, channel detection likelihoods of the utterances are compared. The results are listed in the third line. The rates in this line are obtained for 245 test trials (59 speakers × 5 channels). As observed in the table, GMM and cohort based methods perform comparably. Fusion of the two methods improves the detection performance. We achieve approximately 4-6% absolute improvement in recognition rate with the fusion.

Table 6.6. Percent channel recognition rates.

	GMM-based	Cohort-based	Fusion
Auth. Utt.	73.92	75.08	79.50
Enroll. Utt.	74.01	74.56	79.18
Enroll. Session	76.73	76.33	80.41

We investigate the effects of employing channel detection on speaker verification performance for the best performing handset dependent normalization, namely HC-norm.

EERs are listed in Table 6.7 for the three channel detection methods. DET curves are shown in Figure 6.7. In the table and figure, the ideal performance is also provided in which we assume that channel types are known (or a perfect channel detection system is available). The error rate in this column (i.e., 0.72) can be viewed as a theoretical upper performance bound.

Table 6.7. Effects of employing channel detection to identify enrollment/authentication channels in HC-norm.

	HC-norm (Ideal)	HC-norm (GMM-based)	HC-norm (Cohort-based)	HC-norm (Fusion)
EER%	0.72	0.80	0.76	0.74

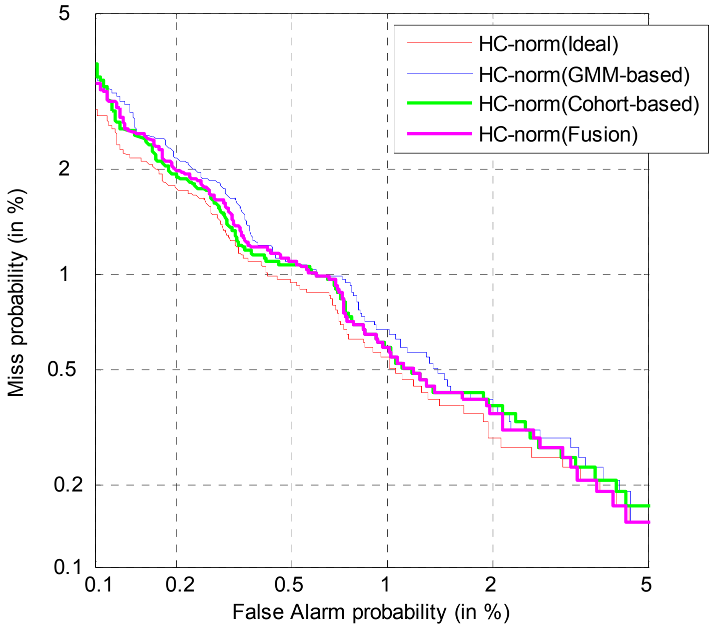

Figure 6.7. Effects of employing channel detection to identify enrollment/authentication channels in HC-norm.

As observed in Figure 6.7 and Table 6.7, performance is slightly degraded when channel detection is employed to identify the channel types. This result may be attributed to the nature of channel detection errors. We may say that most of the errors are due to confusion among channels with similar acoustic characteristics, and therefore do not degrade the performance significantly. In terms of speaker verification accuracy, cohort-based channel detection outperforms GMM-based method. Fusion of the two methods gives the closest error rate to the ideal channel known case.

6.2.2. Spectral Mean Division (SMD)

In this subsection, we test SMD against the baseline CMS using the TDSU-D. We also investigate the case where SMD and CMS are combined together in order to observe whether they provide complementary information to each other. SMD and SMD + CMS are implemented as explained in Section 3.1.2.

DET curves for CMS, SMD and SMD + CMS systems are shown in Figure 6.8, Figure 6.9 and Figure 6.10 for the cases without score normalization, T-norm and HT-norm, respectively. In the first two figures, we can observe that SMD outperforms CMS especially in the high FA region. However, in the low FA region CMS shows slightly better performance. Moreover, we note that using a combination of SMD and CMS results in the best performance. One explanation for this result might be that CMS is focusing more on the normalization of the coarse structure of the spectrum while SMD is also attempting to normalize for fine details. Therefore they offer complementary information and hence increase overall accuracy of the system. On the other hand, only in HT-norm case CMS slightly outperforms both SMD and SMD + CMS methods.

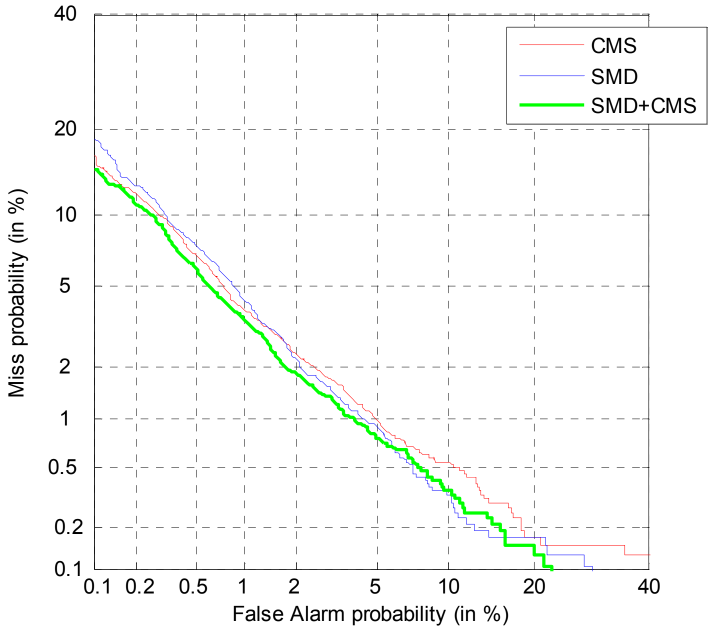

Figure 6.8. DET curves using CMS, SMD and SMD + CMS without score normalization.

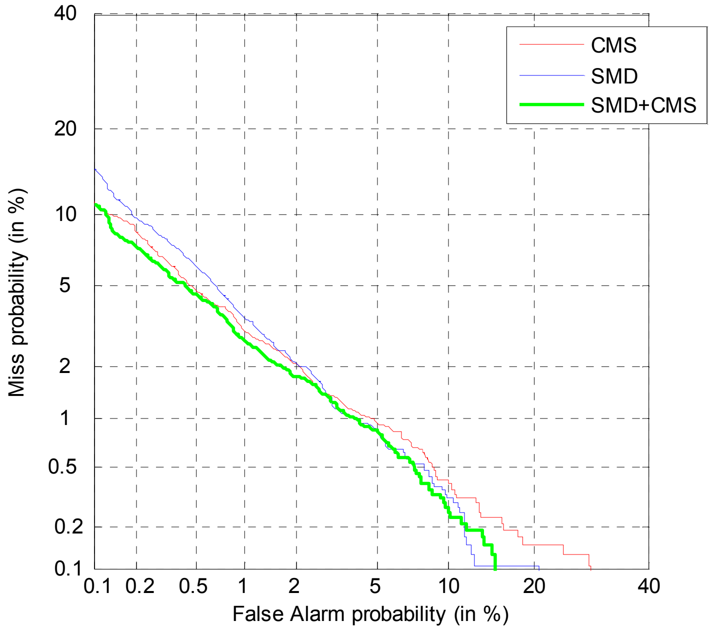

Figure 6.9. DET curves using CMS, SMD and SMD + CMS with T-norm.

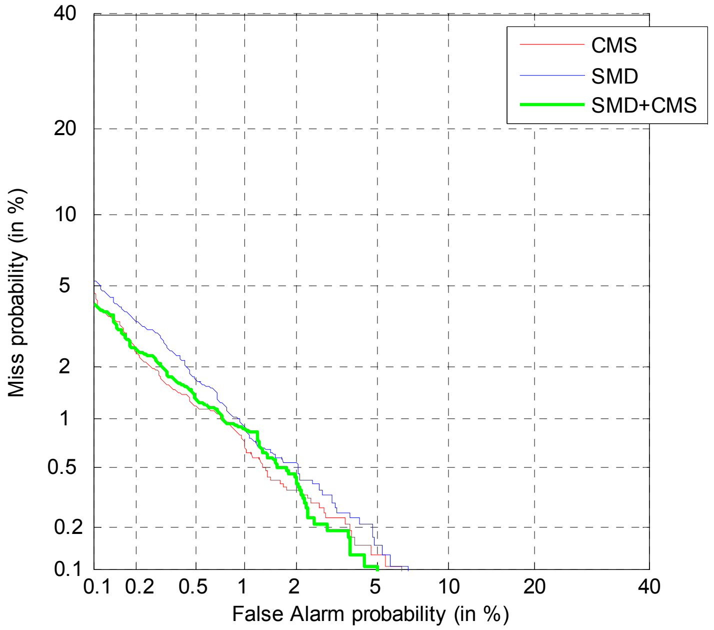

Figure 6.10. DET curves using CMS, SMD and SMD + CMS with HT-norm.

EERs for the three compensation techniques are summarized in Table 6.8. As observed in the table, SMD + CMS yields the best performance for no score normalization and T-norm cases. With the combination of the two methods, the relative reduction in EER is 15.5% for no score normalization and 9.4% for T-norm cases when compared to the baseline CMS experiment. Only in HT-norm, these observations do not hold and CMS performs slightly better than SMD + CMS.

Table 6.8. EER% using CMS, SMD and SMD + CMS.

	CMS	SMD	SMD + CMS
No-norm	2.19	2.04	1.85
T-norm	2.02	2.02	1.83
HT-norm	0.86	0.95	0.90

6.2.3. Prosodic features

Various studies in speaker verification literature reported that high-level features provide complementary information to the low-level spectral features although they do not yield high accuracy separately. For example in (Shriberg et al., 2005; Dehak et al., 2007; Leung et al., 2008; Ferrer et al., 2010) 10-15% relative reduction in EERs is achieved when prosodic systems are combined with the state-of-art spectral systems in text-independent applications. Moreover, these features are known to be less susceptible to channel variations and background noise.

In this subsection, we investigate possible use of prosodic features in our TDSU task. The details of the prosodic systems were discussed in Section 3.1.5. Briefly, we use the sentence HMM method as the baseline spectral system and extract prosodic statistics using the time alignment information obtained from the HMM states. As mentioned earlier speaker models are adapted from speaker independent model using three utterances from the first session in the baseline spectral system. Prosodic statistics are also extracted from the same enrollment utterances. The scores of the prosodic and spectral systems are fused with a three layer neural network. In the following paragraphs, we first describe the system fusion and then present our experimental results.

6.2.3.1. System Fusion. The scores from the spectral and prosodic systems are combined with a three-layer perceptron network. The numbers of neurons in the layers are four, three and one, respectively. Transfer functions for the first two layers are hyperbolic tangent sigmoid. We use linear transfer function in the last layer. In this subsection, we use the same experimental setup as in Section 6.1.5. However, one of the sets is used to train the neural network parameters. This set contains nine test speakers. Verification experiments are conducted with the remaining five sets. Total number of genuine and imposter trials in the five sets are 3885 and 29180, respectively.

6.2.3.2. Verification Results. In Table 6.9, EERs of spectral and prosodic systems are presented for match and mismatch conditions. In the table, scores are normalized with T-norm. In Table 6.10, EERs of spectral, prosodic and fusion systems are given for mixed condition in which match and mismatch trials are accumulated. In the fusion, we combine T-norm (or HT-norm) normalized scores of spectral, duration and pitch features. We did not use other features in the fusion since they did not provide additional improvement.

Table 6.9. EER% for spectral and prosodic systems in match and mismatch conditions. Speaker scores are normalized with T-norm.

	Match	Mismatch
Spectral	0.26	1.93
Duration	6.31	11.42
Pitch	13.51	16.41
Delta-pitch	28.19	32.66
Voiced-unvoiced decision	32.30	38.19
Energy	11.84	30.44

Table 6.10. EER% for spectral, prosodic and fusion systems in mixed condition.

	T-norm	HT-norm
Spectral	2.19	0.98
Duration	10.81	9.55
Pitch	15.98	14.98
Delta-pitch	31.76	30.76
Voiced-unvoiced decision	37.22	36.63
Energy	28.93	25.92
Fusion (Spectral + Pitch + Duration)	1.96	0.88

We can make several observations from Tables 6.9 and 6.10. First, as observed in Table 6.9, mismatch conditions result in higher relative performance degradation in the spectral system when compared to prosodic systems. Among the prosodic features, energy is more susceptible to mismatch conditions. This result might be attributed to the differences in microphone and background noise levels in the recording sessions. Second, from Table 6.10, we observe that HT-norm improves the verification accuracy significantly in the spectral system while the rate of improvement is marginal in all prosodic systems. The first two observations indicate that prosodic features are more robust to channel variations as expected. Third, prosodic features do not yield high accuracy when they are individually employed. Among the prosodic features, the best performance is obtained for duration. Raw pitch value also seems to be a good feature. The error rates in delta-pitch, voiced-unvoiced decision and energy features are much higher when compared to duration and pitch. Fourth, fusion of spectral and prosodic systems improves the verification performance. Absolute reduction in EER is higher in T-norm when compared to HT-norm. Relative reduction in the error rates is approximately 10% for both normalization methods. Although prosodic features are found to be more robust to channel mismatch conditions, they do not provide higher relative improvement in handset-independent normalization.

In Figure 6.11, DET curves for spectral and fusion systems are depicted where the scores are normalized with T-norm. In Figure 6.12, DET curves for HT-norm scores are presented. As observed in the figures, fusion systems outperform spectral systems in almost all operating points of the DET curves. These curves also show that prosodic features might provide complementary information to the spectral features in a TDSU task.

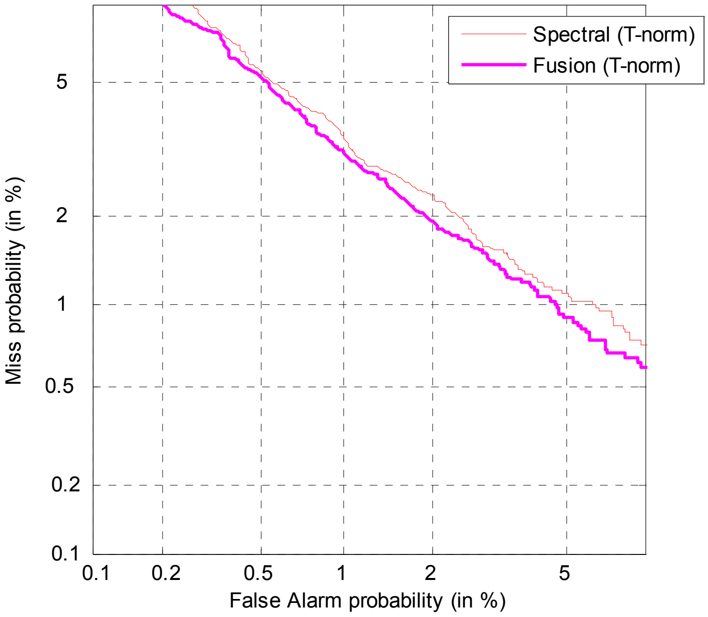

Figure 6.11. DET curves for spectral (T-norm) and fusion (T-norm) systems.

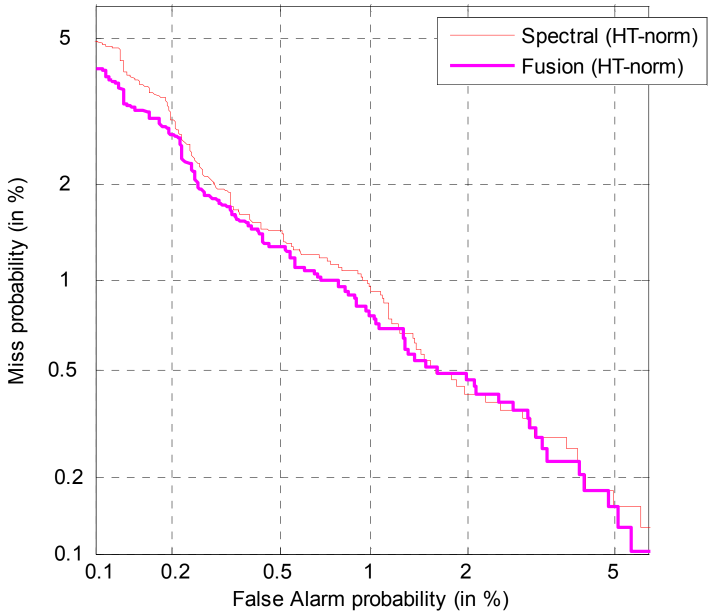

Figure 6.12. DET curves for spectral (HT-norm) and fusion (HT-norm) systems.

7. CONCLUSIONS

In this thesis, we studied model selection and channel variability issues for telephone-based text-dependent speaker verification applications. For this purpose, we collected two multi-channel databases which consist of digit strings and short phrases in Turkish language. In the conclusion, we will summarize our key observations and contributions.

In Chapter 5, we presented our experimental results on the text-dependent variable text (TDVT) database. First of all, we observed that the amount of enrollment and authentication speech has a significant impact on verification performance. We obtained the best performances in multi-channel enrollment case. Moreover, the performance in place of birth was much lower than the other authentication utterances. Second, we realized that context match between enrollment and authentication lexicons plays a crucial role in the performance. HMM outperformed GMM when there exists a context match. When there is no match as in text-independent applications, GMM also performs as well as HMM. Additionally, we can say that in speaker verification tasks with limited data, recording exactly the same utterance in both enrollment and authentication sessions seems to be the best option in terms of verification performance. Third, we found out that increasing enrollment data amount is not always beneficial when there is no context match with the authentication material. Mismatch conditions also degraded the performance in our experiments. Especially severe mismatch cases such as landline and GSM channels showed significant degradation. In order to compensate the effects of channel mismatch, we studied feature domain LTAS filtering, model domain MLLR transformation and score domain H-rank methods together with the widely used CMS. The best performance was obtained when MLLR transformation and H-rank are used with CMS. LTAS filtering added incremental improvement over CMS. Moreover, we observed that employing channel detection prior to the compensation does not degrade the performance significantly when there is a similar channel in the database.

Inspired from Chapter 5’s results, we devoted Chapter 6 to text-dependent single utterance (TDSU) task. In model selection experiments in Chapter 6, sentence HMM outperformed both monophone HMM and GMM based methods. We think that this is mainly because sentence HMM captures co-articulation information better. In order to compensate for channel mismatch conditions, we investigated T-norm, Z-norm, TZ-norm, ZT-norm and C-norm score normalization techniques. The proposed combination of T-norm and Z-norm, namely C-norm, yielded the best performance in the experiments. Additionally, providing handset-channel information to the score normalization process was found to be beneficial. In score normalization section, we also investigated the effects of employing channel detection on verification performance. For channel detection, we proposed a cohort-based procedure in addition to the GMM-based method. Experimental results showed that fusion of the GMM and cohort based methods improves both channel detection and speaker verification accuracy. In the second set of channel mismatch compensation experiments, we tested feature domain SMD method against the well-known CMS. In the experiments, combination of SMD and CMS outperformed any individual method when there is no channel information in score normalization. In the last set of experiments, we combined spectral and prosodic (pitch and duration) features together in order to improve the verification accuracy in the TDSU task. We observed that although the prosodic features individually do not yield high performance, they provide complementary information to the spectral features. We achieved approximately 10% relative reduction in EER when the scores from different systems are fused with a multi-layer neural network.

All the results in this thesis are reported for five handset-channel conditions from approximately fifty speakers. However, there are significantly more handset-channel combinations and recording conditions in practice. In order to observe the influence of different handset types, fifth channel is selected to be a random handset for GSM channel in the TDVT database. The results in Chapter 5 indicate that channel compensation techniques improve the performance in this test condition, too. However, in the future larger text-dependent databases can be collected in order to simulate more realistic scenarios.

APPENDIX A: ERROR RATES FOR THE CLOSED-SET SPEAKER IDENTIFICATION EXPERIMENT IN SECTION 5.1.3

Table A.1. FR% rates in HMM based method for password in Session 1.

	1	2	3	4	5	Average
1	0.00	0.00	1.92	15.69	11.54	5.79
2	5.77	0.00	0.00	11.76	13.46	6.18
3	9.62	0.00	0.00	15.69	19.23	8.88
4	23.08	9.62	9.62	1.96	3.85	9.65
5	25.00	11.54	9.62	11.76	0.00	11.58
6	0.00	0.00	0.00	1.96	0.00	0.39
7	0.00	0.00	0.00	7.84	5.77	2.70

Table A.2. FR% rates in HMM based method for password in Session 2.

	1	2	3	4	5	Average
1	2.13	2.13	6.38	12.77	8.51	6.38
2	6.38	0.00	0.00	4.26	10.64	4.26
3	8.51	0.00	2.13	6.38	10.64	5.53
4	21.28	17.02	10.64	4.26	19.15	14.47
5	27.66	17.02	23.40	21.28	2.13	18.30
6	0.00	0.00	0.00	2.13	0.00	0.43
7	2.13	0.00	0.00	2.13	4.26	1.70

Table A.3. FR% rates in HMM based method for random number in Session 1.

	1	2	3	4	5	Average
1	0.00	1.92	1.92	15.38	15.38	6.92
2	0.00	0.00	0.00	7.69	13.46	4.23
3	3.85	1.92	0.00	3.85	9.62	3.85
4	23.08	9.62	15.38	3.85	13.46	13.08
5	17.31	15.38	15.38	7.69	1.92	11.54
6	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	3.85	7.69	2.31

Table A.4. FR% rates in HMM based method for random number in Session 2.

	1	2	3	4	5	Average
1	8.51	6.38	4.26	14.89	19.15	10.64
2	10.64	2.13	0.00	8.51	10.64	6.38
3	12.77	2.13	0.00	12.77	12.77	8.09
4	19.15	21.28	10.64	4.26	6.38	12.34
5	25.53	19.15	21.28	21.28	6.38	18.72
6	4.26	2.13	0.00	0.00	0.00	1.28
7	2.13	2.13	0.00	4.26	8.51	3.40

Table A.5. FR% rates in HMM based method for repeated phrase in Session 1.

	1	2	3	4	5	Average
1	0.00	1.92	3.85	9.62	9.62	5.00
2	0.00	0.00	0.00	3.85	3.85	1.54
3	1.92	0.00	0.00	1.92	1.92	1.15
4	3.85	1.92	1.92	0.00	1.92	1.92
5	11.54	9.62	11.54	3.85	0.00	7.31
6	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	1.92	3.85	1.15

Table A.6. FR% rates in HMM based method for repeated phrase in Session 2.

	1	2	3	4	5	Average
1	0.00	2.13	0.00	6.38	6.38	2.98
2	0.00	0.00	0.00	2.13	2.13	0.85
3	0.00	0.00	0.00	8.51	4.26	2.55
4	4.26	4.26	2.13	4.26	4.26	3.83
5	14.89	10.64	10.64	10.64	2.13	9.79
6	0.00	0.00	0.00	2.13	0.00	0.43
7	0.00	0.00	0.00	0.00	2.13	0.43

Table A.7. FR% rates in HMM based method for place of birth in Session 1.

	1	2	3	4	5	Average
1	21.15	30.77	38.46	57.69	61.54	41.92
2	34.62	11.54	21.15	55.77	46.15	33.85
3	40.38	26.92	19.23	53.85	53.85	38.85
4	51.92	51.92	48.08	46.15	51.92	50.00
5	59.62	53.85	50.00	51.92	32.69	49.62
6	26.92	17.31	21.15	38.46	38.46	28.46
7	26.92	11.54	17.31	51.92	48.08	31.15

Table A.8. FR% rates in HMM based method for place of birth in Session 2.

	1	2	3	4	5	Average
1	38.30	36.17	29.79	61.70	59.57	45.11
2	38.30	34.04	29.79	55.32	51.06	41.70
3	31.91	34.04	25.53	57.45	59.57	41.70
4	55.32	53.19	59.57	53.19	51.06	54.47
5	53.19	48.94	51.06	61.70	34.04	49.79
6	31.91	38.30	31.91	51.06	38.30	38.30
7	27.66	21.28	19.15	51.06	48.94	33.62

REFERENCES

Adami, A. G., R. Mihaescu, D. A. Reynolds, and J. J. Godfrey, 2003, "Modeling prosodic dynamics for speaker recognition", *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2003 (ICASSP 2003).*

Ariyaeeinia, A. M., J. Fortuna, P. Sivakumaran, and A. Malegaonkar, 2006, "Verification effectiveness in open-set speaker identification", *IEE Vision, Image and Signal Processing,* vol. 153, no. 5, pp. 618-624.

Aronowitz, H., D. Irony, and D. Burshtein, 2005, "Modeling intra-speaker variability for speaker recognition", *Proc. of the Ninth European Conference on Speech Communication and Technology 2005 (INTERSPEECH 2005).*

Auckenthaler, R., M. Carey, and H. Lloyd-Thomas, 2000, "Score normalization for text-independent speaker verification systems", *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42-54.

Avendano, C., and H. Hermansky, 1997, "On the effects of short-term spectrum smoothing in channel normalization", *IEEE Transactions on Speech and Audio Processing,* vol. 5, no. 4, pp. 372-374.

Barras, C., and J. L. Gauvain, 2003, "Feature and score normalization for speaker verification of cellular data", *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2003 (ICASSP 2003)*, vol. 2, pp. 49-52.

BCC, 2010, *BCC Research Report, Biometrics: Technologies and Global Markets*, www.bccresearch.com/report/biometrics-technologies-markets-ift042c.html, 2011.

Beaufays, F., and M. Weintraub, 1997, "Model transformation for robust speaker recognition from telephone data", *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 1997 (ICASSP 1997)*, vol. 2, pp. 1063-1066.

Beigi, H. S. M., S. H. Maes, J. S. Sorensen, and U. V. Chaudhari, 1999, “A hierarchical approach to large-scale speaker recognition”, *Proc. of the Sixth European Conference on Speech Communication and Technology 1999 (EUROSPEECH 1999)*, pp. 2203-2206.

Benzeghiba, M. F., and H. Bourlard, 2003, “Hybrid HMM/ANN and GMM combination for user-customized password speaker verification”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2003 (ICASSP 2003)*, vol. 2, pp. II - 225-8.

Benzeghiba, M. F., and H. Bourlard, 2006, “User-customized password speaker verification using multiple reference and background models”, *Speech Communication*, vol. 48, no. 9, pp. 1200-1213.

Bimbot, F., M. Blomberg, L. Boves, D. Genoud, H. P. Hutter, C. Jaboulet, J. Koolwaaij, J. Lindberg, and J. B. Pierrot, 2000, “An overview of the CAVE project research activities in speaker verification”, *Speech Communication*, vol. 31, no. 2-3, pp. 155-180.

Blouet, R., C. Mokbel, H. Mokbel, E. S. Soto, G. Chollet, and H. Greige, 2004, “Becars: A free software for speaker verification”, *Proc. of the Speaker and Language Recognition Workshop 2004 (Odyssey 2004)*, pp. 145-148.

Boll, S. F., 1979, “Suppression of acoustic noise in speech using spectral subtraction”, *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 27, no. 2, pp. 113-120.

Bolt, R. H., F. S. Cooper, E. E. David, P. B. Denes, J. M. Pickett, and K. N. Stevens, 1970, “Speaker identification by speech spectrograms: A scientists’ view of its reliability for legal purposes”*, Journal of the Acoustical Society of America,* vol. 47, no. 2B, pp. 597-612.

Camlikaya, E., B. Yanikoglu, and H. Erdogan, 2007, “Saklı Markov modeli kullanarak ses ile metin bağımlı kimlik doğrulama”, *Proc. of the IEEE Fifteenth Signal Processing and Communication Applications Conference 2007 (SIU 2007).*

Campbell, J. P., 1995, “Testing with the YOHO CD-ROM voice verification corpus”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 1995 (ICASSP 1995),* vol. 1, pp. 341-344.

Campbell, W. M., D. E. Sturim, and D. A. Reynolds, 2006, “Support vector machines using GMM supervectors for speaker verification”, *IEEE Signal Processing Letters,* vol. 13, no. 5, pp. 308-311.

Cetingul, H. E., E. Erzin, Y. Yemez, and A. M. Tekalp, 2006, “Multimodal speaker/speech recognition using lip motion, lip texture and audio”, *Signal Processing,* vol. 86, no. 12, pp. 3549-3558.

Charlet, D., D. Jouvet, and O. Collin, 2000, “An alternative normalization scheme in HMM-based text-dependent speaker verification”, *Speech Communication,* vol. 31, no. 2-3, pp. 113-120.

Che, C., Q. Lin, and D. S. Yuk, 1996, “An HMM approach to text-prompted speaker verification”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 1996 (ICASSP 1996)*, vol. 2, pp. 673-676.

Chollet, G., J. L. Cochard, A. Constantinescu, C. Jaboulet, and P. Langlais, 1996, “Swiss French Polyphone and PolyVar: telephone speech databases to model inter- and intra-speaker variability”, IDIAP Research Report, IDIAP-RR96-01, 1996.

Das, A., G. Chittaranjan, and G. K. Anumanchipalli, 2008, “Usefulness of text-conditioning and a new database for text-dependent speaker recognition research”, *Proc. of the Ninth Annual Conference of the International Speech Communication Association 2008 (INTERSPEECH 2008)*, pp. 1925-1928.

Das, A. and M. Tapaswi, 2010, “Direct modeling of spoken passwords for text-dependent speaker recognition by compressed time-feature representations”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2010 (ICASSP 2010).*

Davis, S. B., and P. Mermelstein, 1980, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”, *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366.

Dehak, N., P. Dumouchel, and P. Kenny, 2007, “Modeling prosodic features with joint factor analysis for speaker verification”, *IEEE Transactions on Audio, Speech and Language Processing,* vol. 15, no. 7, pp. 2095-2103.

de Luis-Garcia, R., C. Alberola-Lopez, O. Aghzout, and J. Ruiz-Alzola, 2003, “Biometric identification systems”, *Signal Processing*, vol. 83, no. 12, pp. 2539-2557.

Dong, C., Y. Dong, J. Li, and H. Wang, 2008, “Support vector machines based text dependent speaker verification using HMM supervectors”, *Proc. of the Speaker and Language Recognition Workshop 2008 (ODYSSEY 2008).*

Falavigna, D., 1995, “Comparison of different HMM based methods for speaker verification”, *Proc. of the Fourth European Conference on Speech Communication and Technology 1995 (EUROSPEECH 1995)*, pp. 371-374.

Ferrer, L., N. Scheffer, and E. Shriberg, 2010, “A comparison of approaches for modeling prosodic features in speaker recognition”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2010 (ICASSP 2010).*

Furui, S., 1981, “Cepstral analysis technique for automatic speaker verification”, *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254-272.

Furui, S., 2009, “Selected topics from 40 years of research on speech and speaker recognition”, *Proc. of the Tenth Annual Conference of the International Speech Communication Association 2009 (INTERSPEECH 2009).*

Glaeser, A., and F. Bimbot, 1998, “Steps toward the integration of speaker recognition in real-world telecom applications”, *Proc. of the Fifth International Conference on Spoken Language Processing 1998 (ICSLP 1998)*, vol. 4, pp. 1603-1607.

Gravier, G., J. Kharroubi, and G. Chollet, 2000, “On the use of prior knowledge in normalization schemes for speaker verification”, *Digital Signal Processing,* vol. 10, no. 1-3, pp. 213-225.

Gupta, H., V. Hautamaki, T. Kinnunen, and P. Franti, 2005, “Field evaluation of text-dependent speaker recognition in an access control application”, *Proc. of the Tenth International Conference on Speech and Computer 2005*, pp. 551-554.

Gutman, D., and Y. Bistritz, 2002, “Speaker verification using phoneme-adapted Gaussian mixture models”, *Proc. of the Eleventh European Signal Processing Conference 2002 (EUSIPCO 2002)*, vol. 3, pp. 85-88.

Hasan, M. K., S. Salahuddin, and M. R. Khan, 2004, “A modified a priori SNR for speech enhancement using spectral subtraction rules”, *IEEE Signal Processing Letters*, vol. 11, no. 4.

Hardt, D., and K. Fellbaum, 1997, “Spectral subtraction and rasta-filtering in text-dependent HMM-based speaker verification”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 1997 (ICASSP 1997)*, vol. 2, pp. 867-870.

Hebert, M., and D. Boies, 2005, “T-norm for text-dependent commercial speaker verification applications: effect of lexical mismatch”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2005 (ICASSP 2005)*, vol. 1, pp. 729-732.

Heck, L. P., and M. Weintraub, 1997, “Handset dependent background models for robust text-independent speaker recognition”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 1997 (ICASSP 1997)*, vol. 2, pp. 1071-1074.

Hermansky, H., 1990, “Perceptual linear predictive (PLP) analysis of speech”, *Journal of the Acoustical Society of America,* vol. 87, no. 4, pp. 1738–1752.

Hermansky, H., 1992, “RASTA-PLP speech analysis technique”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 1992 (ICASSP 1992)*, pp. 1.121-1.124.

Higgins, A., L. Bahler, and J. Porter, 1991, “Speaker verification using randomized phrase prompting”, *Digital Signal Processing,* vol. 1, no. 2, pp. 89-106.

Huang, W. Y., and B. D. Rao, 1995, “Channel and noise compensation for text dependent speaker verification over telephone”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 1995 (ICASSP 1995)*, vol. 1, pp. 337-340.

Itakura, F., 1975, “Minimum prediction residual principle applied to speech recognition”, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 67-72.

Isobe, T., and J. Takahashi, 1999, “A new cohort normalization using local acoustic information for speaker verification”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 1999 (ICASSP 1999)*, vol. 2, pp. 841-844.

Jin, Q., 2007, *Robust Speaker Recognitio*n, Ph.D. Thesis, Carnegie Mellon University.

Mirghafori, N. and M. Hebert, 2004, “Parameterization of the score threshold for a text-dependent adaptive speaker verification system”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2004 (ICASSP 2004).*

Kajarekar, S. S., N. Scheffer, M. Graciarena, E. Shriberg, A. Stolcke, L. Ferrer, and T. Bocklet, 2008, “The SRI NIST 2008 speaker recognition evaluation system”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2009 (ICASSP 2009)*, pp. 4205 – 4208.

Kenny, P., 2005, “Joint factor analysis of speaker and session variability: Theory and algorithms”, Technical Report CRIM-06/08-13.

Kenny, P., G. Boulianne, P. Ouellet, and P. Dumouchel, 2007, “Joint factor analysis versus eigenchannels in speaker recognition”, *IEEE Transactions on Audio, Speech and Language Processing,* vol. 15, no. 4, pp. 1435-1447.

Kersta, L. G., 1962, “Voiceprint identification”, *Nature*, vol. 196, pp 1253-1257.

Kinnunen, T., and H. Li, 2010, “An overview of text-independent speaker recognition: From features to supervectors”, *Speech Communication*, vol. 52, no. 1, pp. 12-40.

Klusacek, D., J. Navratil, D. A. Reynolds, and J. Campbell, 2003, “Conditional pronunciation modeling in speaker detection”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2003 (ICASSP 2003).*

Koenig, W., H. K. Dunn and L. Y. Lacy, 1946, “The sound spectrograph”, *Journal of the Acoustical Society of America*, vol. 18, pp. 19-49.

Lamel, L. F., and J. L. Gauvain, 2000, “Speaker verification over the telephone”, *Speech Communication*, vol. 31, no. (2-3), pp. 141–154.

LDC, 2011, *Linguistic Data Consortium*, http://www.ldc.upenn.edu/, 2011.

Leggetter, C. J., and P. C. Woodland, 1995, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models”, *Computer Speech and Language,* vol. 9, no. 2, pp. 171-185.

Leung, C. C., M. Ferras, C. Barras, and J. L. Gauvain, 2008, “Comparing prosodic models for speaker recognition”, *Proc. of the Ninth Annual Conference of the International Speech Communication Association 2008 (INTERSPEECH 2008)*, pp. 1945-1948.

Li, K. P., and J. E. Porter, 1988, “Normalizations and selection of speech segments for speaker recognition scoring”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 1988 (ICASSP 1988)*, vol.1, pp. 595-598.

Li, J., Y. Dong, C. Dong, and H. Wang, 2007 “Score normalization technique for text-prompted speaker verification with Chinese digits,” *Lecture Notes in Computer Science,* vol. 4682, pp. 1082-1089.

Li, H., B. Ma, K. A. Lee, H. Sun, D. Zhu, K. C. Sim, C. You, R. Tong, I. Karkkainen, C. L. Huang, V. Pervouchine, W. Guo, Y. Li, L. Dai, M. Nosratighods, T. Tharmarajah, J. Epps, E. Ambikairajah, E. S. Chng, T. Schultz, and Q. Jin, 2009, ”The I4U system in NIST 2008 speaker recognition evaluation”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2009 (ICASSP 2009)*, pp. 4201 – 4204.

Mak, M. W., and S. Y. Kung, 2002, “Combining stochastic feature transformation and handset identification for telephone-based speaker verification”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2002 (ICASSP 2002)*, vol.1, pp. I701-I704.

Mak, M. W., C. L. Tsang, and S. Y. Kung, 2004, “Stochastic feature transformation with divergence-based out-of-handset rejection for robust speaker verification”, *EUROSIP Journal on Applied Signal Processing*, vol. 2004, no. 4, pp. 452-465.

Mak, M. W., R. Hsiao, B. Mak, 2006, “A comparison of various adaptation methods for speaker verification with limited enrollment data”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2006 (ICASSP 2006).*

Makhoul, J., 1975, “Linear prediction: A tutorial review”, *Proceeding of the IEEE*, vol. 63, no. 4, pp. 561–580.

Martin, A., G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, 1997, “The DET curve in assessment of detection task performance”, *Proc. of the Fifth European Conference on Speech Communication and Technology 1997 (INTERSPEECH 1997)*, pp. 1895-1898.

Matsui, T., and S. Furui, 1993 “Concatenated phoneme models for text-variable speaker recognition”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 1993 (ICASSP 1993)*, vol. 2, pp. 391-394.

Matsui, T., and S. Furui, 1994a, “Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM’s”, *IEEE Transactions on Speech and Audio Processing,* vol. 2, no. 3, pp. 456-459.

Matsui, T., and S. Furui, 1994b, “Speaker adaptation of tied-mixture-based phoneme models for text-prompted speaker recognition”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 1994 (ICASSP 1994)*, vol.1, pp 125-128.

Mokbel, C., 2001, “Online adaptation of HMMs to real-life conditions: a unified framework”, *IEEE Transactions on Speech and Audio Processing, vol. 9, no. 4, pp. 342-357.*

Murthy, H. A., F. Beaufays, L. P. Heck, and M. Weintraub, 1999, “Robust text-independent speaker identification over telephone channels”, *IEEE Transactions on Speech and Audio Processing,* vol. 7, no. 5, pp. 554-568.

Nealand, J. H., J. W. Pelecanos, R. D. Zilca, and G. N. Ramaswamy, 2005, “Study of the relative importance of temporal characteristics in text-dependent and text-constrained speaker verification”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2005 (ICASSP 2005)*, vol.1, pp. 653-656.

Neumeyer, L. G., V. V. Digalakis, and M. Weintraub, 1994, “Training issues and channel equalization techniques for the construction of telephone acoustic models using a high-quality speech corpus”, *IEEE Transactions on Speech and Audio Processing,* vol. 2, no. 4, pp. 590-597.

Newman, M., L. Gillick, Y. Ito, D. McAllaster, and B. Peskin, 1996 “Speaker verification through large vocabulary continuous speech recognition”, *Proc. of the Fourth*

International Conference on Spoken Language Processing 1996 (ICSLP 1996), pp. 2419-2422.

NIST, 2011, *National Institute of Standards and Technology. Speaker Recognition Evaluation*, http://www.nist.gov/speech/tests/spk.

Okamoto, H., S. Tsuge, A. Abdelwahab, M. Nishida, Y. Horiuchi, and S. Kuroiwa, 2009, "Text-Independent speaker verification using rank threshold in large number of speaker models", *Proc. of the Tenth Annual Conference of the International Speech Communication Association 2009 (INTERSPEECH 2009)*, pp. 2367-2370.

Orman, O. D., and L. M. Arslan, 2001, "Frequency analysis of speaker identification", *Proc. of The Speaker Recognition Workshop 2001 (ODYSSEY 2001)*, pp. 219-222.

Ortega-Garcia, J., and J. Gonzalez-Rodriguez, 1997, "Providing single and multi-channel acoustical robustness to speaker identification systems", *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 1997 (ICASSP 1997)*, vol. 2, pp. 1107-110.

Panda, A., N. Tripathi, and T. Srikanthan, 2007, "Improved spectral subtraction technique for text-independent speaker verification", *Proc. of the Fifteenth International Conference on Digital Signal Processing 2007*, pp. 595-598.

Pelecanos, J., and S. Sridharan, 2001, "Feature warping for robust speaker verification", *Proc. of the Speaker Recognition Workshop 2001 (ODYSSEY 2001)*, pp. 213-218.

Poza, F., and D. R. Begault, 2005, "Voice identification and elimination using aural-spectrographic protocols", *Audio Engineering Society Twenty Sixth International Conference 2005,* pp. 21-28.

Rabiner, L. R., and B. W. Huang, 1993, *Fundamentals of Speech Recognition,* Prentice Hall Inc., Englewood Cliffs, New Jersey.

Ramasubramanian, V., A. Das, and V. P. Kumar, 2006, “Text-dependent speaker recognition using one-pass dynamic programming algorithm”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2006 (ICASSP 2006).*

Ramos-Castro, D., J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, 2007, “Speaker verification using speaker and test-dependent fast score normalization”, *Pattern Recognition Letters*, vol. 28, no. 1, pp. 90-98.

Reynolds, D. A., 1994, “Experimental evaluation of features for robust speaker identification”, *IEEE Transactions on Speech and Audio Processing,* vol. 2, no.4, pp. 639-643, 1994.

Reynolds, D. A., and R. C. Rose, 1995, “Robust text-independent speaker identification using Gaussian mixture speaker models”, *IEEE Transactions on Speech and Audio Processing,* vol. 3, no. 1, pp. 72-83.

Reynolds, D. A., M. A. Zissman, T. F. Quatieri, G. C. O'Leary, and B. A. Carlson, 1995, “The effects of telephone transmission degradations on speaker recognition performance”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 1995)*, pp. 329-332.

Reynolds, D. A., 1996, “The effects of handset variability on speaker recognition performance: Experiments on the Switchboard corpus”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 1996 (ICASSP 1996)*, vol. 1, pp. 113-116.

Reynolds, D. A., T. F. Quatieri, and R. B. Dunn, 2000, “Speaker verification using adapted Gaussian mixture models”, *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41.

Reynolds, D. A., 2002, “An overview of automatic speaker recognition technology”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2002 (ICASSP 2002)*, vol. 4, pp. 4072-4075.

Reynolds, D. A., 2003, “Channel robust speaker verification via feature mapping”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2003 (ICASSP 2003)*, vol. 2, pp. II-53-6.

Reynolds, D. A., W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, 2003, “The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2003 (ICASSP 2003).*

Reynolds, D. A., and W. Campbell, 2007, *Text-independent speaker recognition*, Springer Handbook of Speech Processing and Communication, Springer-Verlag GMBH, Heidelberg, Germany.

Reynolds, D. A*.*, 2008, *Gaussian mixture models*, Encyclopedia of Biometric Recognition, Springer.

RNCOS, 2011, *RNCOS Market Research Report, Global Biometric Forecast to 2012*, http://www.rncos.com/Report/IM140.htm, 2011.

Rose, R. C., J. Fitzmaurice, E. M. Hofstetter, and D. A. Reynolds, 1991, “Robust speaker identification in noisy environments using noise adaptive speaker models”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 1991 (ICASSP 1991)*, vol. 1, pp. 401-404.

Rosenberg, A. E., and S. Parthasarathy, 1996, “Speaker background models for connected digit password speaker verification”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 1996 (ICASSP 1996)*, pp. 81-84.

Rosenberg, A. E., O. Siohan, and S. Parthasarathy, 2000 “Small group speaker identification with common password phrases”, *Speech Communication*, vol. 31, no. 2-3, pp. 131-140.

Shahin, I., 2008, "Speaker identification in the shouted environment using suprasegmental hidden Markov models", *Signal Processing*, vol. 88, no. 11, pp. 2700-2708.

Shriberg, E., L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, 2005, "Modeling prosodic feature sequences for speaker recognition", *Speech Communication*, vol. 46, no. 3-4, pp. 455-472.

Shriberg, E., 2007, "Higher-level features in speaker recognition", *Lecture Notes in Computer Science,* vol. 4343, pp. 241-259.

Solomonoff, A., W. M. Campbell, and I. Boardman, 2005, "Advances in channel compensation for SVM speaker recognition", *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2005 (ICASSP 2005)*, pp. 629-632.

Sphinx, 2009, *Carnegie Mellon University Sphinx, Open Source Toolkit for Speech Recognition*, http://cmusphinx.sourceforge.net/, 2009.

Sturim, D.E., D. A. Reynolds, R. B. Dunn, T. F. Quatieri, 2002, "Speaker verification using text-constrained Gaussian mixture models", *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2002 (ICASSP 2002),* vol. 1, pp. 677-680.

Sturim, D. E., and D.A Reynolds, 2005, "Speaker adaptive cohort selection for Tnorm in text-independent speaker verification", *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2005 (ICASSP 2005)*, vol.1, pp. 741-744.

Sturim, D. E., W. Campbell, Z. Karam, D. A. Reynolds, and F. Richardson, 2009, "The MIT Lincoln Laboratory 2008 speaker recognition system", *Proc. of the Tenth Annual Conference of the International Speech Communication Association 2009 (INTERSPEECH 2009).*

Subramanya, A., Z. Zhang, A. C. Surendran, P. Nguyen, M. Narasimhan and A. Acero, 2007, "A generative-discriminative framework using ensemble methods for text-dependent speaker verification", *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2007 (ICASSP 2007),* vol. 4, pp. IV-225 - IV-228.

Tadj, C., P. Dumouchel, M. Mihoubi, and P. Ouellet, 1999, "Environment adaptation and long term parameters in speaker identification", *Proc. of the Sixth European Conference on Speech Communication and Technology 1999 (EUROSPEECH 1999)*, pp. 1015-1018.

Talkin, D., 1995, "A robust algorithm for pitch tracking (RAPT)", in Speech Coding and Synthesis edited by W. B. Kleijn, and K. K. Paliwal, Elsevier, New York, pp. 495–518.

Teunen, R., B. Shahshahani, and L. P. Heck, 2000, "A model-based transformational approach to robust speaker recognition", *Proc. of the Sixth International Conference on Spoken Language Processing 2000 (ICSLP 2000)*, vol. 2, pp. 495-498.

Toledano, D. T., C. Esteve-Elizalde, J. Gonzalez-Rodriguez, R. F. Pozo, and L. H. Gomez, 2008, "Phoneme and sub-phoneme T-normalization for text-dependent speaker recognition", *Proc. of the Speaker and Language Recognition Workshop 2008 (ODYSSEY 2008).*

Tosi, O., H. J. Oyer, W. Lashbrook, C. Pedney, J.Nichol, and W. Nash, 1972, "Experiment on voice identification", *Journal of the Acoustical Society of America,* vol. 51, no. 6B, pp. 2030-2043.

van Vuuren, S., 1996, "Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch", *Proc. of the Fourth International Conference on Spoken Language Processing 1996 (ICSLP 1996)*, pp. 1788-1791.

Vogt, R., B. Baker, and S. Sridharan, 2005, "Modeling session variability in text-independent speaker verification", *Proc. of the Ninth European Conference on Speech Communication and Technology 2005 (INTERSPEECH 2005).*

Weber F., L. Manganaro, B. Peskin, and E. Shriberg, 2002, ”Using prosodic and lexical information for speaker identification”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2002 (ICASSP 2002).*

Wu, W., T. F. Zheng, and M. Xu, 2006, “Cohort-based speaker model synthesis for channel robust speaker recognition”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2006 (ICASSP 2006)*, vol. 1, pp. I-893~896.

Xiang, B., U. V. Chaudhari, J. Navratil, G. N. Ramaswamy, and R. A. Gopinath, 2002 “Short-time Gaussianization for robust speaker verification”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2002 (ICASSP 2002)*, vol. 1, pp. 681-684.

Yamada, M., M. Sugiyama, and T. Matsui, 2010, “Semi-supervised speaker identification under covariate shift”, *Signal Processing,* vol. 90, no. 8, pp. 2353-2361.

Yegnanarayana, B., S. R. M. Prasanna, J. M. Zachariah, C. S. Gupta, 2005, “Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system”, *IEEE Transactions on Speech and Audio Processing,* vol. 13, no. 4, pp. 575-582.

Yiu, K. K., M. W. Mak, and S. Y. Kung, 2007, “Environment adaptation for robust speaker verification by cascading maximum likelihood linear regression and reinforced learning”, *Computer Speech and Language*, vol. 21, no. 2, pp. 231-246.

Yoma, N. B., and T. F. Pegoraro, 2002, “Robust speaker verification with state duration modeling”, *Speech Communication,* vol. 38, no. 1-2, pp. 77-88.

Yoma, N. B., C. Garreton, C. Molina, and F. Huenupan, 2008, “Unsupervised intra-speaker variability compensation based on Gestalt and model adaptation in speaker verification with telephone speech”, *Speech Communication,* vol. 50, no. 11-12, pp. 953-964.

Young, S., G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, 2006, *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering Department.

Yuo, K. H., T. H. Hwang, and H. C. Wang, 2005, “Combination of autocorrelation-based features and projection measure technique for speaker identification”, *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 565-574.

Yu, K., J. Mason, and J. Oglesby, 1995, “Speaker recognition using hidden Markov models, dynamic time warping and vector quantization”, *IEE Proceedings on Vision, Image and Signal Processing*, vol. 142, no. 5, pp. 313-318, 1995.

Zhu, X., B. Millar, I. Macleod, M. Wagner, F. Chen, and S. Ran, 1994, “A comparative study of mixture-Gaussian VQ, ergodic HMMs and left-to-right HMMs for speaker recognition”, *International Symposium on Speech, Image Processing and Neural Networks*, vol. 2, pp. 618-621.

Zhu, D., B. Ma, H. Li, and Q. Huo, 2007, “A generalized feature transformation approach for channel robust speaker verification”, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2007 (ICASSP 2007)*, vol. 4, pp. 61-64.

Zheng, R., S. Zhang, and B. Xu, 2005, “A comparative study of feature and score normalization for speaker verification”, *Lecture Notes in Computer Science,* vol. 3832, pp. 531-538.