Identity Verification Using Voice and its Use in a Privacy Preserving System

by

Eren Çamlıkaya

Submitted to the Graduate School of Engineering and Natural Sciences in partial fulfillment of
the requirements for the degree of
Master of Science

Sabancı University
August 2008

Identity Verification Using Voice and its Use in a Privacy Preserving System

APPROVED BY:

Associate Prof. Dr. Berrin Yanıkoğlu (Thesis Supervisor)	_____
Assistant Prof. Dr. Hakan Erdoğan (Thesis Co-Supervisor)	_____
Associate Prof. Dr. Albert Levi	_____
Associate Prof. Dr. Erkay Savaş	_____
Assistant Prof. Dr. Gözde Ünal	_____

DATE OF APPROVAL: _____

ACKNOWLEDGEMENTS

As humans, we always need guidance and encouragement to fulfill our goals. I would like to thank my thesis supervisor Assoc. Prof. Dr. Berrin Yanıkoğlu for her endless understanding and care throughout this study and also for providing the opportunity, the motivation and the resources for this research to be done. I also owe many thanks to Assist. Prof. Dr. Hakan Erdoğan for his kindness and precious help as my co-supervisor.

I am also grateful to Assoc. Prof. Dr. Erkay Savaş, Assoc. Prof. Dr. Albert Levi and Assist. Prof. Dr. Gözde Ünal for their participation in my thesis committee and their comprehensive reviews on my thesis. Moreover, I would like to offer my special thanks to Prof. Dr. Aytül Erçil and all members of the VPAlab to provide such a fruitful atmosphere of research and friendship with their support throughout this thesis. I also would like to thank "The Scientific and Technological Council of Turkey (TUBITAK - BIDEB)" very much for their support via the scholarship with code "2210" throughout my graduate education. Without their help, I would not be able to complete my research and earn my degree.

Finally, I offer thanks to my dearest friend Dila Betil for her love and motivation during my years at Sabancı University. Lastly, I would like to thank to my mother, Ferhan Özben and my sister, Ayşe Çamlıkaya for raising me and encouraging me to follow my own decisions at all times.

August 2008 Eren Çamlıkaya

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Since security has been a growing concern in recent years, the field of biometrics has gained popularity and became an active research area. Beside new identity authentication and recognition methods, protection against theft of biometric data and potential privacy loss are current directions in biometric systems research.

Biometric traits which are used for verification can be grouped into two: physical and behavioral traits. Physical traits such as fingerprints and iris patterns are characteristics that do not undergo major changes over time. On the other hand, behavioral traits such as voice, signature, and gait are more variable; they are therefore more suitable to lower security applications. Behavioral traits such as voice and signature also have the advantage of being able to generate numerous different biometric templates of the same modality (e.g. different pass-phrases or signatures), in order to provide cancelability of the biometric template and to prevent cross-matching of different databases.

In this thesis, we present three new biometric verification systems based mainly on voice modality. First, we propose a text-dependent (TD) system where acoustic features are extracted from individual frames of the utterances, after they are aligned via phonetic HMMs. Data from 163 speakers from the TIDIGITS database are employed for this work and the best equal error rate (EER) is reported as 0.49% for 6-digit user passwords.

Second, a text-independent (TI) speaker verification method is implemented inspired by the feature extraction method utilized for our text-dependent system. Our proposed TI system depends on creating speaker specific phoneme codebooks. Once phoneme codebooks are created on the enrollment stage using HMM alignment and segmentation to extract discriminative user information, test utterances are verified by calculating the total dissimilarity/distance to the claimed codebook. For benchmarking, a GMM-based TI system is implemented as a baseline. The results of the proposed TD system (0.22% EER for 7-digit passwords) is superior compared to the GMM-based system (0.31% EER for 7-digit sequences) whereas the proposed TI system yields worse results (5.79% EER for 7-digit sequences) using the data of 163 people from the TIDIGITS database .

Finally, we introduce a new implementation of the multi-biometric template framework of Yanikoglu and Kholmatov [12], using fingerprint and voice modalities. In this framework, two biometric data are fused at the template level to create a multi-biometric template, in order to increase template security and privacy. The current work aims to also provide cancelability by exploiting the behavioral aspect of the voice modality.

ÖZET

Güvenlik konusu günümüzde giderek artan bir endişe olduğundan biyometrik araştırmalar daha da önem kazanmıştır. Biyometrik konusundaki güncel çalışmalar yeni kimlik doğrulama tekniklerinin yanı sıra, biyometrik verilerin hırsızlığına ve bu verilerden veya verilerin tutulduğu veritabanlarından kişisel bilgilerin ortaya çıkarılmasına karşı önlemler üzerine yoğunlaşmıştır.

Kimlik tanımlama için kullanılan biyometrik özellikler iki gruba ayrılabilir: fiziksel ve davranışsal özellikler. Parmakizi ve iris örüntüleri gibi fiziksel özellikler zaman içerisinde çok fazla değişmeyen özelliklerdir. Öte yandan ses, imza, yürüyüş gibi davranışsal özellikler daha değişken bir yapıda olup aşırı güvenlik gerektirmeyen sistemler için daha uygundurlar. Ses ve imza gibi davranışsal özellikler diğer biyometrik özelliklere nazaran söylenilen kelimenin ya da atılan imzanın değişmesi ile aynı özelliği kullanarak farklı şablonlar oluşturabilme avantajına sahiptirler. Farklı uygulamalarda farklı şablonlar kullanılması, veritabanların karşılaştırılarak kullanıcı hakkında bilgi çıkarılmasını önleyebilecek önemli bir etkendir.

Bu tez kapsamında, ses kullanarak üç farklı biyometrik sistem sunulmuştur. İlk olarak fonem bazlı Saklı Markov Modeller (SSM) yardımıyla hizalanmış işitsel parolalardan akustik öznitelikler çıkarılarak metin bağımlı bir sistem önerilmiştir. TIDIGITS veritabanına ait 163 kişinin 6 haneli işitsel parolaları kullanılmış ve en iyi Eşit Hata Oranı (EHO) %0.49 olarak hesaplanmıştır.

İkinci olarak, bir önceki bölümde anlatılan öznitelik çıkarma yönteminden esinlenerek tasarlanmış bir metin bağımsız konuşmacı tanıma sistemi gerçeklenmiştir. Önerilen bu metin bağımsız sistem konuşmacı fonem çizelgelerinden faydalanmaktadır. Eğitim için kullanılan işitsel parolaların fonem bazlı SMM yardımıyla hizalanıp konuşmacılar arasındaki farkı en fazla gözetebilecek özniteliklerin çıkarılması ile her konuşmacı için fonem çizelgeleri hazırlanmıştır. Sınama aşamasında ise sınanacak işitsel parola ile iddia edilen kişinin fonem çizelgesi arasındaki toplam uzaklığa bakılmaktadır. Karşılaştırma için Karma-Gaus-Modelleri (KGM) kullanan bir metin bağımsız konuşmacı tanıma sistemi gerçeklenmiştir. Önerilen metin bağımlı sistemin sonuçları (7 haneli işitsel parolalar için %0.22 EHO) KGM tabanlı sisteme (7 haneli işitsel parolalar için %0.31 EHO) göre daha iyidir. Öte yandan TIDIGITS veritabanına ait 163 konuşmacının bilgileriyle oluşturulup önerilen metin bağımsız

sistem (7 haneli işitsel parolalar için %5.79 EHO) karşılaştırılan diğer iki sisteme göre kötü performans sergilemiştir.

Son olarak Yanıkoğlu ve Kholmatov [12] tarafından önerilen çoklu biyometrik şablon çerçevesine ses ve parmakizi kullanarak yeni bir örnek sunulmuştur. Bu çerçevede iki biyometrik veri şablon seviyesinde birleştirilerek şablon güvenliği ve mahremiyetinin artırılması amaçlanır. Bu çalışmada davranışsal bir biyometrik olan ses verilerinin kullanılması ile, var olan çerçeveye şablonun iptal edilebilmesi özelliği eklenmiştir.

CHAPTER 1

INTRODUCTION

Due to increasing security concerns, person identification and verification have gained significance over the last years. Identification or verification of a claimed identity can be based on 3 major themes: "what you have", "what you know" or "who you are". Historically, the first two themes have been the main methods of authentication. Electronic identification cards are also commonly used as tokens in entering secure areas. Similarly, a credit card and its pin number form a simple example of the fusion of the two themes. Systems that are based on the theme of "who you are" are classified as biometric systems. Biometric systems utilizes pre-recorded physical (e.g. iris, fingerprint and hand shape) or behavioral (e.g. signature, voice and gait) traits of a person for later authentication.

The main characteristics distinguishing different biometric modalities include universality (whether everyone has that trait); measurability (whether that biometric can be easily measured); stability (whether the trait changes significantly over time); forgeability (whether someone else can easily forge your biometric trait); and whether the biometric can be changed at will (whether the person can change his own trait to hide his identity). In these regards, physical traits stand out as they are quite universal, mostly stable, hard to forge and changeable at will. Some other physiological biometrics and most behavioral biometrics are more varying, either due to ageing or other reasons such as stress. However they may be better suited for a particular security application (e.g. online banking over the phone).

A major concern with the use of biometric technologies is the fear that they can be used to track people if biometric databases are misused. A related concern is that once compromised, a physiological biometric (e.g. fingerprint) cannot be canceled (one cannot get a new fingerprint). The privacy and cancelability concerns lead researches to find new solutions, often combining cryptography and biometrics in recent years [35, 2, 18, 12].

In this thesis, we present biometric authentication systems based mainly on voice modality. Voice has certain advantages over other biometrics, in particular acceptability of its use for identity authentication, as well as its suitability for certain

tasks such as telephone banking. A further advantage of voice is that it provides a cancelable biometric within text-dependent speaker verification systems.

In voice verification systems, different levels of information can be extracted from a speech sample of a user. As summarized by Day and Nandi [41], lexical and syntactic features of voice such as language and sentence construction are at the highest level. These features are highly dependent on the spoken text, however very costly computationally. In order to extract high level features, automatic speech recognition tools need to be utilized first. After extracting the words uttered in a given speech sample, lexical or syntactic analysis can be done as described by Day and Nandi [41]. This means, in order to extract high-level features, additional calculations are needed after lower level features are extracted first. At the lower levels, there are prosodic features like intonation, stress and rhythm of speech. These features also depend on the spoken text, but also on how the text is uttered. Next, there are the phonetic features based on the sound of the syllables which also vary according to the uttered text. Lastly, low level acoustic features can be extracted to acquire information about the generation of the voice by the speaker and these are considered to be text independent [41].

Verification systems based on voice are divided into two main groups: text-dependent (TD) and text-independent (TI) systems. During enrollment to a text-dependent system, the speaker is asked to repeat a fixed text which is considered to be his/her password. Then, a user specific template or model is constructed from the collected reference samples of the spoken password. In authentication, the utterance is compared with the template of the claimed identity. If the similarity is above a certain predefined threshold, the utterance is accepted and the user is verified. In text-independent systems, mostly low levels of information from spectral analysis is used for identification or verification since higher levels of features are mainly dependent on the text. Thus, TI systems require longer training sessions and varying voice samples to include all sounds for all possible voice combinations to create statistical speaker models whereas a fewer repetitions of the spoken password are enough to create a template or a model in TD systems. For a TI system, the statistical speaker models are created from the phonemes extracted from the collected data. Then, during the testing phase, a voice sample is compared with the text-independent user-specific model and the speaker is verified according to the similarity scores.

An important factor which determines the success of a voice verification system is the duration and the scope of the training and testing sessions. As mentioned above, longer training sessions by using numerous utterances results in better description of the templates in TD systems or speaker models in TI systems. Similarly, longer test utterances provide better verification performance for both TI and TD

systems: the longer the utterance, the more information can be extracted from the voice sample.

While comparing speaker identification or verification systems the database size is also an important factor to evaluate the reliability of the system. Larger databases provide more confidence in the reported results. Therefore, when comparing performances of different systems, one should consider the size of the database used in order to have a fair opinion on the performances of the compared systems. Generally, public databases such as TIDIGITS, YOHO or NIST are used for benchmarking different algorithms in speaker identification or verification.

Beside speaker verification systems (text-dependent in Chapter 3 and text-independent in Chapter 4), we present an implementation of a multi-biometric framework that combines voice and fingerprint, in Chapter 5. Combinations of biometric traits are preferred due the following reasons: their lower error rates, increased privacy and cancelability if one of the biometrics is a behavioral trait like voice. Using multiple biometric modalities has been shown to decrease error rates by providing additional useful information to the classifier. Fusion of any behavioral or physiological traits can occur in various levels. Different features can be used by a single system at the feature, template or decision level [9]. For this work, voice and fingerprint, are fused at the template level and both biometric features are combined to be used by a single verifier. The second gain obtained by combining multiple biometrics at the template level is privacy. In summary, privacy is increased since the combined template do not reveal individual biometrics. Finally, changing spoken password in a text-dependent speaker verification scenario adds cancelability to the combined biometric template.

The remainder of the thesis is as follows. After an introduction to Hidden Markov Models in Chapter 2, we present a new method for text-dependent speaker verification through extraction of fixed-length feature vectors from utterances, in Chapter 3. The system is faster and uses less memory as compared to the conventional HMM-based approach, while having state-the-art results. In our system, we only use a single set of speaker-independent monophone HMM models. This set is used for alignment, whereas for the conventional HMM-based approach, an adapted HMM set for each speaker is constructed in addition to a speaker independent HMM set (also called universal background model in that context). This requires much higher amount of memory as compared to the proposed approach. In addition, during testing only a single HMM alignment is required as compared to two HMM alignments using a universal background model and a speaker model for the conventional approach. Thus, verification is also faster with the approach introduced in this thesis.

In Chapter 4, we propose a text-independent speaker verification system using

phoneme codebooks. These codebooks are generated by aligning the enrollment utterances using phonetic HMMs and creating MFCC-based fixed-length feature vectors to represent each phoneme. Through creating phoneme codebooks, we tried to extract discriminative speaker information at the phoneme level. However the results of this chapter is not in par with state-the-art TI verification results.

In Chapter 5, we introduce a new implementation of the multi-biometric template framework of Yanikoglu and Kholmatov [12], using fingerprint and voice modalities. In this framework, two biometric data are fused at the template level to create a combined, multi-biometric template, in order to increase both security and privacy of the system. In addition to the first implementation of this framework, which used two fingerprints and showed increases in both security and privacy, the implementation presented here also provides cancelability. Cancelability of the multi-biometric template is achieved by changing the pass-phrase uttered by the speaker, since the generated voice minutiae depends on the pass-phrase comprised of a unique sequence of phonemes.

Finally, in the last chapter, our contribution on the literature of speaker verification and multi-biometric template generation is summarized and some possible extensions are given.

CHAPTER 2

Hidden Markov Models

The hidden Markov model is a statistical model used for modeling an underlying Markov process whose states are hidden to the outside, but observable through the associated outcomes. The model consists of a finite set of hidden states, where the state transitions are controlled by the transition probabilities. Furthermore; in any state, there is also a probability distribution for emitting a certain outcome. It is only the outcome or the observations which are visible externally, and the challenge is to predict the state sequence of the process which are "hidden" to the outside; hence the name hidden Markov model.

An HMM can be fully characterized by [48]:

(1) The number of states in the model, N. Most of the time there is some physical significance attached to the set of states of the model even though the they are hidden. The states are denoted by $S = \{S_1, S_2, S_3...S_N\}$ and the state at time t as q_t.

(2) The number of distinct observation symbols per state, M. Observation symbols corresponds to the physical outcome (e.g. LPC or MFCC vectors as speech features) of the system being modeled. The observations can belong to a discrete alphabet or the set of real vectors. In case of MFCC vectors for voice data, the set of possible observations are the set of real vectors. For the case of a discrete alphabet, M is the discrete alphabet size and individual symbols are denoted as $V = \{v_1, v_2, v_3...v_M\}$ for an observation sequence $O = \{O_1, O_2, O_3...O_t\}$. Here, O_1 and O_2 can be the same observation symbol v_k.

(3) The state transition probabilities among hidden states A=$\{a_{ij}\}$ where
$a_{ij} = P(q_{t+1} = S_j | q_t = S_i$ for $1 \leq i, j \leq N$.

(4) The observation probability distribution in state j for the emission of a visible observation v_k, B=$\{b_{jk}\}$ where

$b_{jk} = P(v_k(t)|q_t = S_j)$ for $1 \leq j \leq N$ and $1 \leq k \leq M$ for a discrete alphabet of outcomes.

(5) The initial state distribution $\pi = P(q_1 = S_i)$ for $1 \leq i \leq N$.

In order to generate a hidden Markov model, the parameters described above need to be calculated from training examples. In a Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but observations (e.g. voice feature vectors) affected by the states are. Therefore, observation probabilities need to be calculated as well. At the end, the trained model can be used to estimate the most likely state sequence, or the probability that the observations were generated by that model.

HMMs are widely used in pattern recognition applications such as speech, handwriting, and gesture recognition, as well as bioinformatics. In case of speech recognition, observable parameters would be speech feature vectors (LPC, MFCC, etc) of an incoming utterance and the hidden states would be the associated phonemes.

Figure 2.1 below shows the state transition diagram of a simple HMM. There are only two states S_1 and S_2 and two possible observations V_1 and V_2. As described above, a_{ij} shows the transition probabilities from state i to state j. Moreover, b_{jk} shows observation probabilities for observing outcome k from state j.

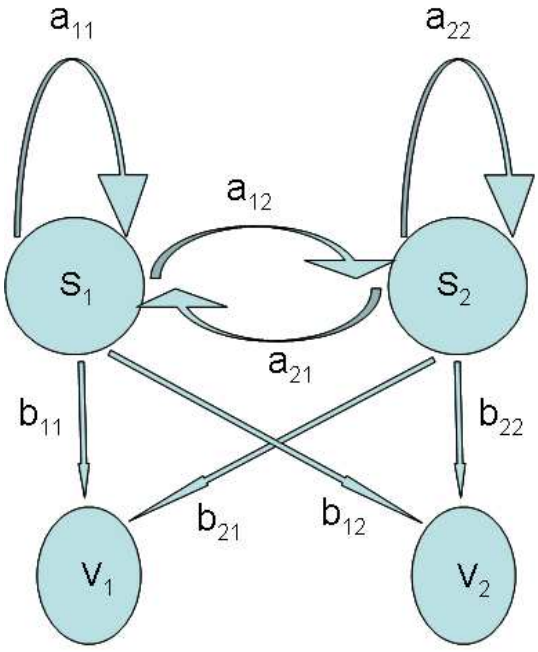

Figure 2.1: States (S_1, S_2) and observations (V_1, V_2) are illustrated by ellipses where the state transition and observation probabilities are illustrated by arrows.

There are three central issues associated with HMMs given the form and parameters described above. The first problem is referred as the likelihood computation problem [45]. It constitutes the computation of the probability of a particular output $P(O|\lambda)$ given the observation sequence $O = \{O_1, O_2, O_3...O_t\}$ and the model $\lambda = (A, B, \pi)$. This problem mainly addresses the case if there are multiple models to choose from for a given observation. An example scenario would be to choose the most likely speaker-dependent model for a set of feature vectors which belong to a test pass-phrase. To solve this problem, forward algorithm [29] or backward algorithm [26] can be employed.

The second problem is referred as the decoding problem [45] and constitutes finding the most likely sequence of hidden states $Q = \{q_1, q_2, q_3...q_t\}$ that could have generated an output sequence $O = \{O_1, O_2, O_3...O_t\}$ given the model $\lambda = (A, B, \pi)$ and the output. The solution to this problem helps to find the corresponding phonemes or words for each feature vector in a given speech sample for a speech recognition scenario. To solve this problem, Viterbi algorithm is employed whose details can be found in [19].

The third and the most difficult problem is referred as the learning problem [45] and constitutes the estimation of model parameters $\lambda = (A, B, \pi)$ to maximize $P(O|\lambda)$ given the observations. In fact, given any finite observation sequence there is no optimal way of estimating these parameters. As a practical solution, iterative approaches such as Baum-Welch or Expectation-Maximization methods are employed to locally maximize $P(O|\lambda)$. This problem constitutes the training session of a word-based or phoneme-based HMM to be employed in a speech recognition system. After the parameters are optimized, the likelihood of a test sequence of feature vectors to a word or phoneme model can be easily calculated. The details of the algorithms can be found in [27, 8]. The details of aforementioned algorithms are not given in this chapter since the theoretical background of these algorithms are beyond the scope of this work.

CHAPTER 3

Text-Dependent Speaker Verification

In this chapter, we present a novel text-dependent speaker verification system. For text-dependent verification systems, features from multiple utterances of the users are compared with the features extracted from the test utterance. Here, temporal information plays an important role since the sequence of extracted features from the utterance determines the decision of verification. For the proposed verification system, acoustic features are extracted from individual frames and utterances are aligned via phonetic HMMs both for enrollment and verification. After the alignment, fixed-length feature vectors are extracted from the utterances depending the uttered text independent of the time it takes to utter that text. For enrollment, every user in the database is assigned a 6-digit password and reference vectors are extracted from utterances of these unique user passwords in order to acquire speaker statistics. For verification, the test vector extracted from the test utterance is fed into a previously trained classifier. Bayesian classifier and a linear classifier in conjunction with Principal Component are used to verify a test utterance for this work.

3.1 Previous Work

Diversity of text-dependent systems mainly arises from the types of extracted voice features and speaker template/model creating schemes. In general, hidden Markov models (HMM) and Dynamic Time Warping (DTW) are used for the alignment of utterances [42, 14]. Bellagarda et al introduced a text-dependent system by using singular value decomposition for spectral content matching and DTW for temporal alignment of the utterances [23]. For every user, 4 utterances were used as reference, 4 utterances were used genuine and 2 utterances from impostors who were given access to the original enrollment pass-phrases of the claimed speaker. In addition to that, impostors tried to mimic genuine users by changing accent and intonation. On a private database of 93 people (48 genuine, 45 impostor), their result show an EER around 4%.

Yegnanarayana employed difference features such as pitch and duration, along with other well known spectral features such as Mel-Frequency Cepstral Coefficients (MFCC) to construct a TD speaker verification system [13]. DTW was used for utterance matching and error rate is reported to be under 5% for a private database of 30 speakers where all users uttered the same text according to the claimed id from a limited pool of sentences.

Ramasubramanian et al proposed an MFCC-based TD speaker verification system where multiple word templates were created for each speaker. For testing, a variable text (sequence of digits) was prompted to speakers on different testing sessions and is aligned via dynamic programming. This means, the text was known to forgers for the proposes text-dependent speaker verification system. The success rate was reported to increase with the number of templates for the same word where the results were given for 1 to 5 templates per digit. In particular, authors found that when using the TIDIGITS database (100 people), the best error rate was under 0.1% when using 5 templates per digit. However, part of this low error rate was due to cohort normalization, which was done by scaling the test utterance score with the best matching impostor in the database. Therefore, the task here was more like identification which increases the success of this closed-set speaker verification system [54].

Subramanya et al proposed a text-dependent system using the likelihood ratio test by comparing global and adapted user-specific HMMs [10]. They obtained user-specific HMMs from global models of digit utterances using discriminative learning approaches. In order to verify a test utterance Subramanya et al. calculated likelihood scores of both global HMMs and user-specific HMMs derived from the global models. Here, the global models act like background models for the speaker verification task. Moreover, a boosting procedure was applied and weighted word-level likelihood scores were fused with utterance level scores. With this approach, some words (digits in this case) had more discriminative power when the likelihood of utterance models were calculated for the scenario that the impostors know the claimed passwords. A portion of the YOHO corpus is used where they have achieved an EER of 0.26% for 6 digit pass-phrases in comparison with the baseline likelihood ratio test method which gives 0.63% EER.

Similarly, Liu et al proposed a system using segmental HMMs [60] whose states were associated with sequences of acoustic feature vectors rather than individual vectors to explore the role of dynamic information in TD systems . They obtained a 44% reduction in false acceptance rate using the segmental model compared with a conventional HMM on the YOHO corpus.

3.2 Proposed Method

Text-dependent speaker verification assumes the existence of a pass-phrase. In fact, often multiple utterances of the same pass-phrase is used to create the user template. In this work, we use pass-phrases that consists of digit sequences. An overview of the proposed verification system is illustrated in Figure 3.1.

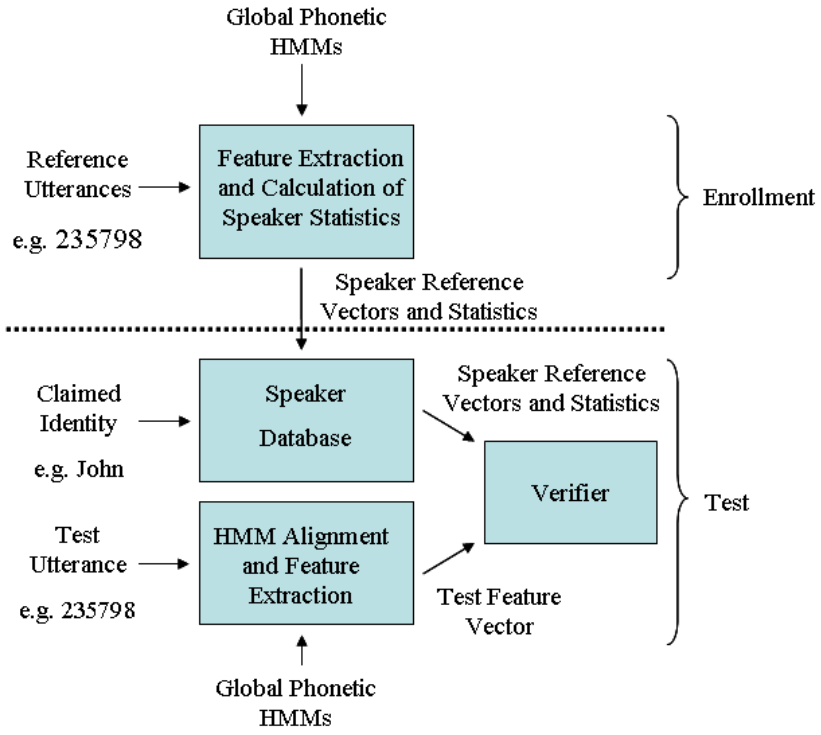

Figure 3.1: System overview: The test utterance is compared with the reference vectors of the claimed identity and accepted if their dissimilarity is low.

For enrollment and verification, the utterances are first aligned using a speaker-independent HMM model of the claimed pass-phrase. Then fixed-length feature vectors are extracted from the aligned utterances and dissimilarities of each reference vector with the test vector is calculated. Using the distance scores to the reference set, a classifier decide whether the the test utterance is genuine or forgery. Details of the verification process are explained in the following subsections.

3.2.1 Feature Extraction

The features employed in speaker recognition systems should successfully be able to define the vocal characteristics of the speaker and distinguish it from the voices of other speakers. Short spectra of speech signals give information about both the spoken words and the voice of the speaker. In particular, we use the Mel frequency cepstral coefficients (MFCCs) features in this work. MFCCs utilize the

logarithmic energies of the speech data after being filtered by nonuniform frequency filters, in a manner similar to the human hearing system. Then, discrete cosine transform is applied to the filtered speech data for further decorrelation of the spectral features [52]. To extract the MFCC features, an utterance is divided into 30ms frames with 10ms overlap and cepstral analysis is applied to each frame. As a result, each 30ms frame is represented by a 12-dimensional vector $< c_1, ..., c_{12} >$ consisting of MFCCs for this work. Beside MFCCs, another approach is to use linear prediction coding (LPC) coefficients or a combination of LPC and MFC coefficients. LPC analysis is based on the linear model of speech production and is very suitable for speech analysis and synthesis purposes. However, we used MFCCs as features to represent utterances since state-of-the-art automatic speaker recognition systems are based on the spectral features [15].

The database used for this work was originally designed for speech recognition and was not suitable for text-dependent speaker verification systems. In order to obtain multiple sequences of the same password, we segmented the utterances in the database into digits, using phonetic HMMs. Thus, the feature extraction in our framework is preceded by the alignment of the utterances (references and query) after extracting MFCC features from individual frames.

The alignment is done using an HMM of the spoken password of the claimed identity. These pass-phrase models are formed by concatenating the HMMs of its constituent phonemes. As an example, the hidden Markov model of the pass-phrase "235798" is formed by concatenating the phoneme models of "t", "uw", "th" for "2" and so on. The goal of aligning pass-phrases is to remove the silence frames and segment the utterance into the phonemes of the pass-phrase. At the end, corresponding frames and phonemes are revealed with silences.

The phoneme models in turn are 3-state, monophone HMMs, constructed for each phoneme found in the digits of the English language. They are speaker-independent models, trained using a separate part of the database. The details of the training process is described in section 2. Phonetic HMMs are commonly used in speech recognition and 3-state monophone models are generally preferred to model phoneme transitions. The alignment process is illustrated in Figure 3.2.

After the alignment, frames which correspond only to the middle state of each phoneme are kept, while the remaining frames (those corresponding to the 1^{st} and 3^{rd} states) are deleted. This is done to use only steady-state portions of each phone, and eliminate the start and end states that are more affected by the neighbouring phones. We then calculate the mean feature vector of cepstral coefficients for each phoneme. After the calculations, each phoneme p is represented by a 12-dimensional mean vector F_p. Using this method, fixed-length feature vectors can be extracted from varying-length utterances consisting of the same digits. The concatenation of the

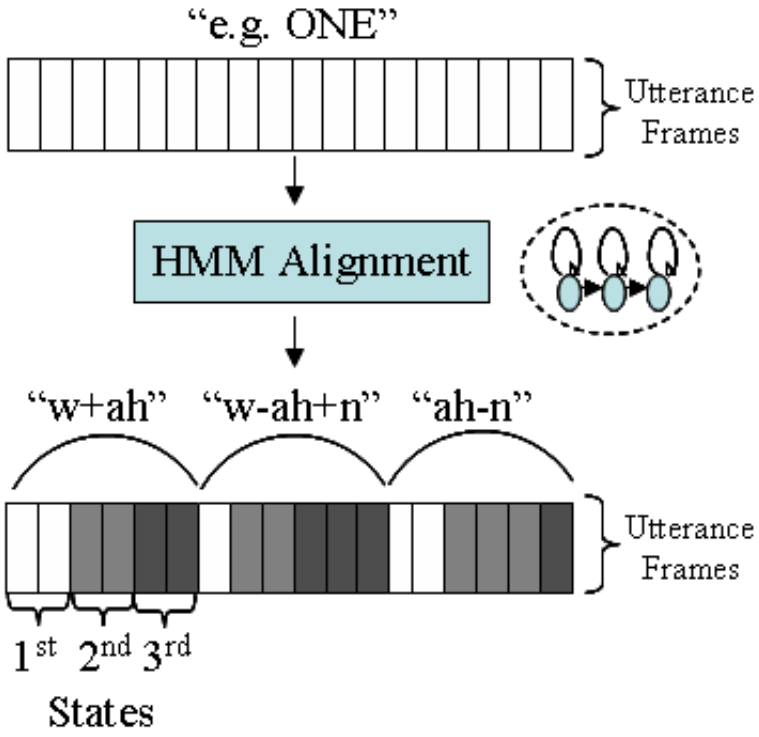

Figure 3.2: Alignment with the global HMM: Previously trained 3-state phonetic HMMs are used in aligning an utterance with the corresponding password model (e.g. "ONE"), which consists of a sequence of the corresponding phoneme models (e.g. "w", "ah", "n").

mean vectors of the phonemes then forms the feature vector for the entire utterance. Creation of fixed-length pass-phrase feature vectors is illustrated in Figure 3.3.

In our experimental setup, test utterances consist 4 or 6 digits (e.g. "2357981"). Since digits in the English language is composed of three phonemes on average, test utterances are composed of around 12 or 18 phonemes. This means, 12 or 18 phoneme vectors are extracted from each test utterance on average.

3.2.2 Enrollment

For speaker verification, it is necessary to go through a speaker specific enrollment session. The purpose of the enrollment session is to create password references for each speaker in the database. First, we randomly selected 4 or 6-digit passwords (e.g. "235798") for each speaker where each digit is used only once in a password to make best use of the available data in the database. Then, artificial password feature vectors are created for each speaker by segmenting and recombining the available utterances in the enrollment set after MFCC based feature vectors are extracted by the feature extraction method described in 3.2.1. We call this reference feature set P_j for each speaker j to be compared during verification tests. This process is illustrated in Figure 3.4.

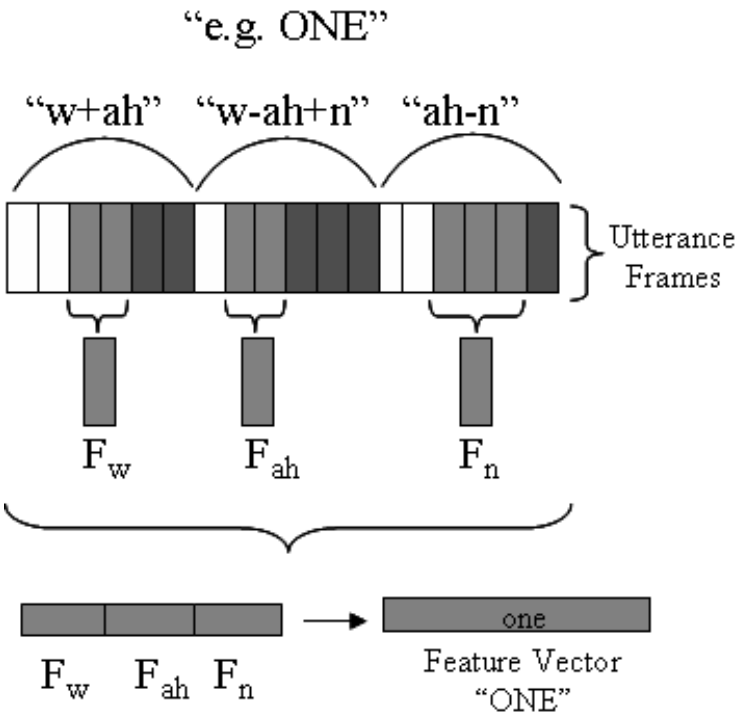

Figure 3.3: Derivation of the feature vector for the whole utterance: First and third phases of the phonemes are discarded and the average of feature vector of the middle phase frames are concatenated to obtain the feature vector.

Later, these reference feature vectors of a speaker are **pairwise** compared. and similarity scores are calculated between each pair of vectors. Hence, if there are N reference vectors for each speaker $N(N-1)$ distances per speaker are calculated. To find the similarity/distance between the feature vectors of two utterances, we used the trimmed Euclidean distance metric [16]. In this metric, the Euclidean distance is measured between two feature vectors after discarding the highest valued dimensions of the difference vector, so that the remaining dimensions are more robust to certain noise artifacts. We have used the same percentage (10%) of discarded dimensions, as in [16]. Note here that the feature vectors are all the same length, regardless of the length of the utterances, due to the feature extraction process.

Using these distances, the following statistics defining the variation among a user's reference templates are extracted, as in [5]:

- average of the nearest neighbor distances,
- average of the farthest neighbor distances,
- minimum of the average distance to all neighbors,
- average distance from reference vectors to the mean vector of the reference feature set P_j

Average of the nearest neighbor distances may indicate how similar a reference utterance to expect, given a query utterance. Average of the farthest neighbor

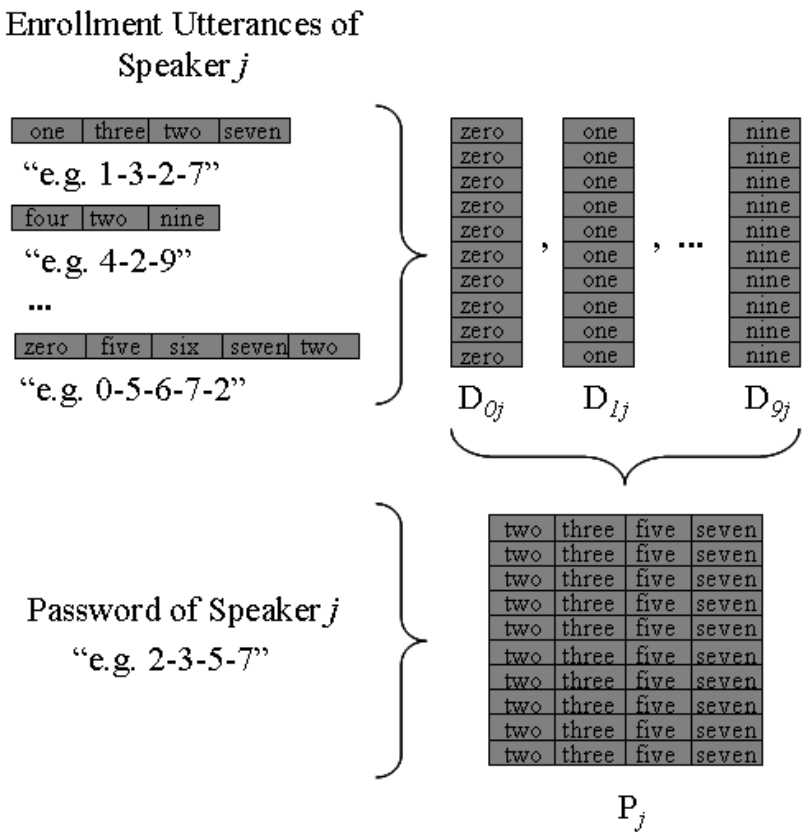

Figure 3.4: The creation of artificial reference passwords: Parsed digit utterances are concatenated to form reference password utterances and feature vectors are extracted.

distances may indicate how far a reference utterance would be at most, given a query utterance. By computing the minimum of the average distance to all neighbors, we in fact designate the **template** utterance which is closest to all other references. User's reference features set P_j together with the calculated parameters are stored to be used in the verification process.

3.2.3 Verification

For verification, the MFCC-based fixed-length feature vector is extracted from the query utterance first. The query feature extraction is done as described in 3.2.1, following the alignment with the corresponding global HMM model. This feature vector is then compared with each reference utterance of the claimed identity, as well as its mean reference vector.

As a result of the comparisons between the query and reference vectors, we find the distances to the closest reference vector, the farthest reference vector, the template reference vector (defined as the one having the smallest total distance to the other reference utterances) and the mean reference vector (the mean of the reference vectors). We use these four distances as input features for the final decision (accept

or reject), after appropriate normalization. In this work, distances are normalized by the corresponding averages of the reference set (e.g. averages of the nearest and farthest neighbor distances of the reference utterances), as described in 3.2.2 and previously used in [5]. Note that normalizing the measured distances eliminates the need for user-dependent thresholds, so the final features were used in comparison to a fixed, speaker-independent threshold. The verification process which is shown by a box in Figure 3.1 is further illustrated in Figure 3.5.

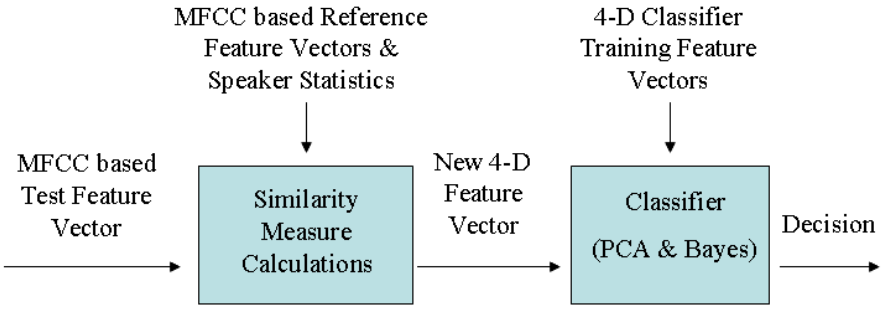

Figure 3.5: Verification process: A 4-dimensional feature vector is extracted from the MFCC-based test vector by calculating the distance to the closest reference vector, the farthest reference vector, the template reference vector and the mean vector of reference vector set of the claimed identity. This 4-dimensional feature vector is later classified as genuine or forgery by a previously trained classifier.

For this work, we have experimented with two different classifiers that take as input the normalized distances and return a decision on the query utterance. The training set used to train the classifiers consists of normalized distances obtained from 163 genuine utterances (1 from each user) and 7472 forgery utterances (1 from each impostor in the same group of each claimed user), not used in the testing, as described in Section 3.3.

One of the classifiers is a Bayes classifier assuming normal distributions for the normalized distances, while the other is a linear classifier following Principal Component Analysis (PCA) of the normalized distances. According to the Bayes theorem, the test utterance should be assigned to the class (genuine or forgery) having the largest posterior probability $\mathrm{P}(C_k|X)$, given the 4-dimensional normalized distance vector, X, in order to minimize the probability of utterance misclassification. These posterior probabilities are calculated as in Eq. 3.1:

$$P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)} \tag{3.1}$$

where k denotes classes (either genuine or forgery).

Using the Bayes classifier, the prior probabilities of classes, $\mathrm{P}(C_g)$ for genuine and $\mathrm{P}(C_f)$ for forgery, are assumed to be equal since we do not know the exact

statistics. Following this assumption, the discriminant function for the Bayesian classifier further simplifies to $g(X) = P(X|C_g) - P(X|C_f)$ and class-conditional probabilities $P(X|C_k)$ needed to determine the decision boundaries were estimated from the training set utterances.

With PCA, 4-dimensional feature vectors (normalized distances) are reduced to 1-dimensional values by projecting them onto the eigenvector (principal component) corresponding to the largest eigenvalue of the training set. Then, a threshold value is found to separate the genuine and forgery utterances of the validation data. This threshold is later used in classifying the test utterances after projecting the 4-dimensional feature vectors onto the same principal component. Details on PCA can be found in [24].

3.3 Database

We used the TIDIGITS database which is originally constructed for speech recognition of digits. The database consists of varying length sequences of digits (e.g. "235798"), uttered by 326 speakers in the database (111 men, 114 women, 50 boys, and 51 girls). Each person utters 77 digit sequences, with length varying between 1 and 7, for a total of 253 digits. Hence, each one of the 11 digits (0-9, and "oh") is uttered roughly 23 times (= 253/11) and at least 16 times by each person. The TIDIGITS database contains a single data collection session whereas some databases may contain multiple sessions distributed over time to form more robust templates.

The database is originally designed for speech recognition and was not suitable for text-dependent speaker verification systems. In order to obtain multiple sequences of the same password, as needed in a text-dependent speaker verification, we segmented the utterances in the database into digits, using the previously described HMMs. This resulted in 16 or more utterances of the same digit by each user. Ten of each of these digits are used for enrollment (reference passwords), 5 for verification (genuine and forgery tests) and 1 for training the classifiers.

Utterances from half of the speakers of each of the 4 groups (men, women, boys, girls) are used to train the phonetic HMMs while the remaining 163 speakers (56 men, 57 women, 25 boys, and 25 girls) are used in constructing the enrollment, genuine and forgery sets described below. In fact, this 163 speaker subset is what we refer to as the "database" throughout the thesis. The utterances from the remaining speakers are used to train the phonetic HMMs.

To create the enrollment set, we randomly picked a 4 or 6-digit password for each user in the database and created artificial utterances of this password by combining segments of the constituent digits. A password here is a string of non-

repeated digits, so as to best use the available data. After creating the enrollment set, genuine test cases were created using the unused digits of the same user (5/16+). As for forgery tests, two sets of tests were constructed according to password blindness. For the password-blind tests (PB), forgery feature vectors are created by the concatenation of **random** digit sequences of the forger (4 or 6-digits, matching the length of the password to be forged) after feature extraction. Password-known tests are conducted to simulate the scenario of a stolen password. Therefore, forgery utterances are created using the **same** digit sequence as the claimed password for the PK tests.

The enrollment set thus consists of a total of 1630 (10 utterances x 163 speakers) reference passwords where each recorded digit is used only once. The genuine test set contains 815 (5 utterances x 163 speakers) genuine passwords constructed from genuine, segmented digit recordings. The forgery test set contains 7472(56x55 for men + 57x56 for women + 25x24 for boys + 25x24 for girls) forgery passwords constructed from segmented digit recordings of other people. Here, each speaker forges every other speaker in the **same** group (men/women/ boys/girls) with only one utterance of the claimed password. In other words, each speaker is forged by all the remaining speakers within the group once who knows his/her password. Since there are only 5 utterances of each digit by each speaker in the verification set, a recorded digit of a speaker is used multiple times for the creation of forgery utterances whereas necessary digits are used only once for creating genuine test utterances. Finally, the training set used for training classifiers contains a set of 163 genuine (constructed from **unused** digit samples of each user) and 7472 forgery utterances (created from **reference** digits of other users that are not used in testing). Thus, genuine or forgery utterances in training set are not used in testing, though they do come from the same speaker set.

One may think that the artificial construction of the passwords does not result in a realistic database with proper coarticulation of the consecutive digits. However, removing the frames corresponding to the first and third states of monophone models (as in our model) reduces the effect of coarticulation since those states are affected by coarticulation the most. Hence, while coarticulation effects would exist with real data (passwords uttered as a digit sequence), we believe that the results would be largely unaffected. Artificial database creation is also used by other authors [54, 10].

3.4 Results

The performance evaluation of both speaker verification systems proposed in this work is done by considering the false acceptance (FAR) and false rejection (FRR) rates during the tests. FAR is calculated as the ratio of falsely verified

impostor utterances to the total number of imposter tests and FRR is calculated as the ratio of falsely rejected genuine utterances to the total number of genuine tests. EER and HER indicate Equal Error rate (where FRR and FAR are made equal by changing the acceptance threshold) the and Half Error Rate (average of FRR and FAR, when EER cannot be obtained).

Separate tests are conducted according to password blindness where the forger did not know the claimed password(PB and PK); the classification method (Bayes and PCA); the password length (4 or 6 digit); and whether the forgers were selected from the same group as the person being forged, or the whole population (same group - SG or all groups - AG). As one can expect, the former (SG) is a more challenging scenario.

Perfect verification results (0% EER) were achieved for the password-blind (PB) scenario, with both classifiers and for both password lengths; therefore, we only list the results for the more challenging password-known case in Table 3.1. The results using PCA-based classifier (shown in bold) are the best, with 0.61% EER for 6-digit passwords for the SG scenario and 0.39% for the AG scenario while the HER rates for the Bayes classifier are higher.

Scenario	Bayes			PCA
	FRR	FAR	HER	EER
6Digit & Same Group (SG)	1.47	0.12	0.80	**0.61**
6Digit & All Groups (AG)	1.47	0.05	0.76	**0.39**
4Digit & Same Group (SG)	1.60	0.62	1.11	**1.10**
4Digit & All Groups (AG)	1.22	0.30	0.76	**0.63**

Table 3.1: False Reject Rate, False Accept Rate, Half Error Rate and Equal Error Rates are given for the password-known scenario and 4 or 6-digit passwords, for different classification methods (Bayes, PCA) and whether the forger was selected from the same group as the person being forged, or the whole population.

The DET figure showing how FAR And FRR changes with different acceptance thresholds is shown in Figure 3.6, for the PCA method.

Further tests were done to see if the performance would improve, if we knew the group (men/women/boys/girls) of the forged person. Note that this information can be derived from a person's age and gender which are public information. For this case, separate classifiers were trained for each group, using as forgers other people from the same group. In other words, if a man was being forged, we used a classifier trained with only information coming from adult male subjects. The results in Table 3.2 show that the error rates for men and women are very similar (0.31 and 0.36%), while that of children (boys and girls groups) are almost twice as much (0.73 and 0.98%). This can be explained as the younger groups showing

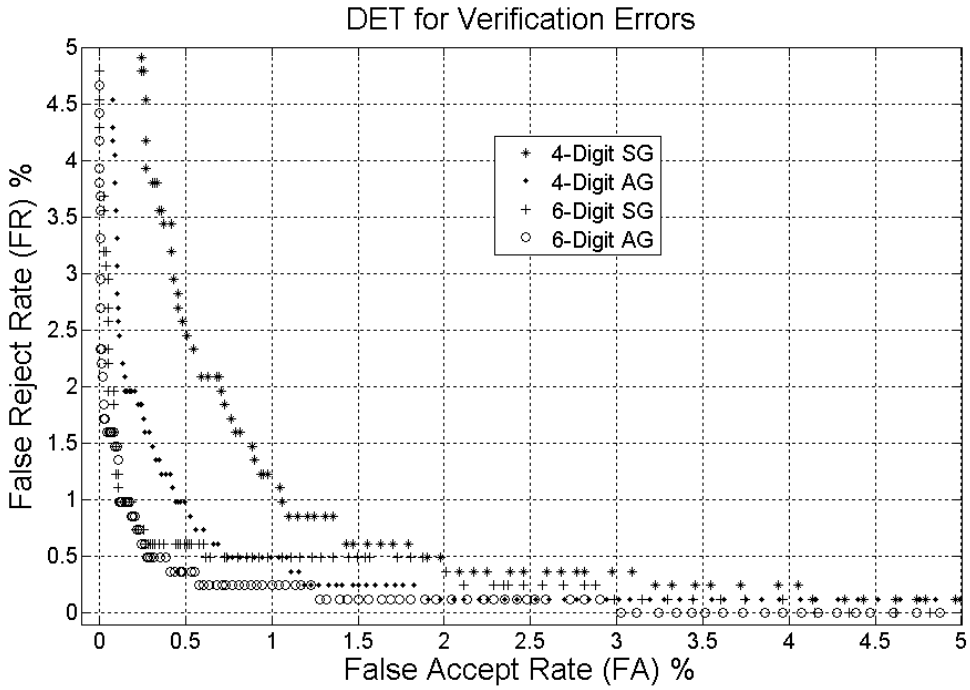

Figure 3.6: DET curves for different password lengths, different forger source, using the password-known scenario and the PCA-based classifier.

more variability in their utterances since the features used in this classification are normalized distances to the reference set of the forged user. Overall, the average EER given in Table 3.2 is slightly lower than the comparable result in Table 3.1 (0.49 vs 0.61%).

Group	Bayes			PCA
	FRR	FAR	HER	EER
women(57)	0.70	0.09	0.40	**0.31**
men(56)	0.71	0.13	0.42	**0.36**
boy(25)	0.80	0.50	0.65	**0.73**
girls(25)	2.40	0.50	1.55	**0.98**
average(163)	0.98	0.23	0.61	**0.49**

Table 3.2: Error rates are given separately for each group, for the password-known scenario, 6-digit passwords, and different classification methods (Bayes, PCA). For these results, the classifiers are trained separately for each group.

3.5 Summary and Contributions

In this chapter, we presented a new method for text-dependent speaker verification. The system is faster and uses less memory as compared to the conventional HMM-based approach. In our system, we only use a single set of speaker-independent monophone HMM models. This set is used for alignment, whereas for

the conventional HMM-based approach, an adapted HMM set for each speaker is constructed in addition to a speaker independent HMM set (also called universal background model in that context). This requires much higher amount of memory as compared to the proposed approach. In addition, during testing only a single HMM alignment is required as compared to two HMM alignments using a universal background model and a speaker model for the conventional approach. Thus, verification is also faster with the approach introduced in this thesis.

The results from our system (0.61 and 0.39% EER) may be compared to the results of Ramasubramanian et al. (under 0.1% EER) who have used the same database under similar conditions [54]. They use multiple utterances of the same digit to create digit templates which are used in verifying utterances of a known digit sequence. However, their EER is lower through cohort normalization using a closed-set verification scenario. In other words, the decision mechanism does not only know about the similarity of the query utterance, but also the similarity of the forgery utterances, which significantly improves verification performance.

Similarly, the results by Subramanya et al [10] who created a database suitable for text-dependent verification from the original YOHO database may also be compared to ours. However, their results of 0.26% should be compared to the average of "men" and "women" groups (0.34%) in our work since "boys" and "girls" groups do not exist in the YOHO database.

CHAPTER 4

Text-Independent Speaker Verification

We have also implemented a text-independent speaker verification method using the TIDIGITS database. Text-independent speaker verification systems are designed to verify any query utterance without the information of the uttered words or sentences. Many TI speaker verification methods have been proposed in literature. These methods mainly differ by their feature selection and speaker modeling processes. Most popular approaches for speaker modeling are Gaussian mixture models (GMM) and support vector machines (SVM), as well as their derivatives and combinations. In addition, other techniques such as vector quantization (VQ) and utterance level scoring have also been used [21, 61]. These issues are discussed thoroughly in [46].

We implemented a text-independent speaker verification method using the TIDIGITS database. Our proposed system depends on creating speaker specific phoneme codebooks. Once phoneme codebooks are created on the enrollment stage using HMM alignment and segmentation, test utterances are verified by calculating the total dissimilarity/distance to the claimed codebook. An overview of the proposed verification system is illustrated in Figure 4.1.

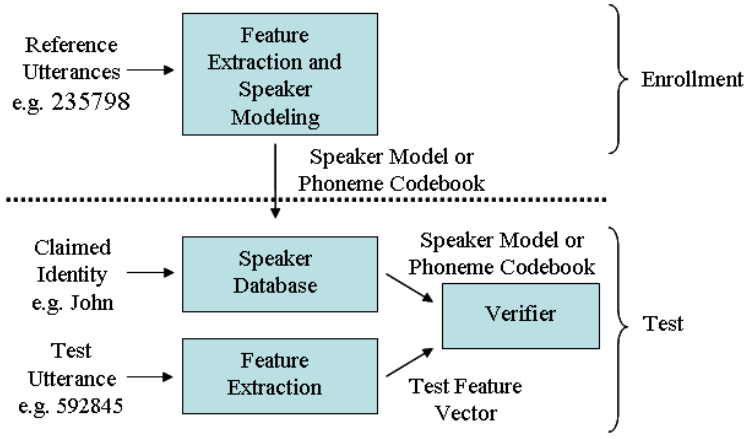

Figure 4.1: System overview: The test utterance is compared with the phoneme codebooks of the claimed identity and accepted if their dissimilarity is low.

For enrollment, we assume that transcriptions of the enrollment utterances are available; in other words, we know the verbal information of the utterances. This way, enrollment utterances are segmented via phoneme-based HMMs and vector codebooks are created. On the other hand, we do not make this assumption in verification, since the task at hand is text-independent verification. Thus, query utterances are not segmented via HMMs for verification since we cannot predict the uttered words or sentences. Although this is possible via an automatic speech recognizer (ASR), the results would not be 100% correct and thus would be misleading for the speaker verifier.

The results for the proposed text-independent speaker verification method is compared with the most popular and successful approach which employs Gaussian Mixture Models to model vocal characteristics of speakers. In order to make an objective comparison with the previously described text-dependent system, new set of tests are conducted using the same amount of enrollment and verification voice data.

4.1 Previous Work

Although cepstral features are employed dominantly in literature, several other features are also investigated. Day and Nandi proposed a TI speaker verification system where different features such as linear prediction coefficients (LPC), perceptual linear prediction coefficients (PLP) and MFCC (acoustic, spectral, etc.) are fused and the speaker verification is done via applying genetic programming methods [41]. Furthermore, the effects of dynamic features such as spectral delta features with novel delta cepstral energies (DCE) on TI speaker verification is investigated by Nostratighods et al [32]. Zheng et al, adopted the GMM-UBM approach for proposing new features derived from the vocal source excitation and the vocal tract system. This new feature is named wavelet octave coefficients of residues (WOCOR) and is based on time-frequency analysis of the linear predictive residual signal [38].

Recently, the GMM employing a universal background model (UBM) with MAP speaker adaptation has become the dominant approach in TI speaker verification. UBM is also a GMM which serves as a background distribution of human acoustic feature space. Current state-of-the art techniques are adopted from this GMM-UBM method by proposing different adaptation and decision criteria to create discriminative speaker models [51, 43, 40]. Several of these adaptation methods are examined in [31]. Moreover, Nuisance attribute projection (NAP) and factor analysis (FA) are also examined to provide improvements over the baseline GMM-UBM method [11].

Beside GMMs, support vector machines are also quite powerful tools that are

used for speaker verification. M.Liu et al proposed a system using SVMs together with the features from the adoption of GMMs [30]. Wan and Renals proposed a system based on sequence discriminant SVM and showed improvement with respect to the well known GMM method [55]. Campbell et al modeled high-level features from frequencies of phoneme n-grams in speaker conversation and fused them with cepstral features to be used by SVMs [57].

4.2 GMM-based Speaker Verification

The GMM framework is a very successful method in the literature of text-independent speaker verification. As a baseline, a GMM-based system is implemented using the TIDIGITS database for this work.

4.2.1 Enrollment

The features employed in a text-independent speaker recognition systems should also be able to define the vocal characteristics of the speaker and distinguish it from the voices of other speakers without employing the temporal information in the uttered text. As described in the previous chapter, short spectra of speech signals gives information about both the spoken words and the voice of the speaker. We also used the Mel frequency cepstral coefficients (MFCCs) features for text-independent speaker verification. First, MFCC features are extracted to represent each frame with a 12-D feature vector for both enrollment and verification as described briefly in section 3.2.1.

For the GMM based method, all frames in the utterances of the enrollment set are used to train the mixture of Gaussians to model the speakers in the database unlike the segmentation process via HMMs. Generally speaking, an N_g component Gaussian mixture for N_d dimensional input vectors has the following form:

$$P(x|M) = \sum_{i=1}^{N_g} a_i \frac{1}{(2\pi)^{N_d/2}|\Sigma_i|^{1/2}} \times exp\left(-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right) \tag{4.1}$$

where P($x|M$) is the likelihood of an input vector $\mathbf{x}$ given the Gaussian mixture model M. N_d equals 12 for our case since we extract 12-D MFCC features from individual frames. The mixture model consists of a weighted sum over N_g Gaussian densities, each parametrized by a mean vector μ_i and a covariance matrix Σ_i where a_i are the mixture weights. These weights are constrained to be non-negative and sum up to one. Since the acoustic space is limited with digits in the English language for this work, the number of mixtures, N_g, is chosen as 32. The parameters of a Gaussian mixture model a_i, μ_i and Σ_i for i=1...N_g are estimated using the maximum likelihood criterion and the EM (expectation maximisation) algorithm [8]. All frames

of utterances in the enrollment set are employed for this estimation process regardless of the uttered text. Although it is usual to employ GMMs consisting of components with diagonal covariance matrices, we employ GMMs with full covariance matrix for better modeling.

For the GMM based method, each speaker is represented by a unique speaker model consisting of density weights a_i, density mean vectors μ_i, and covariance matrices Σ_i where i = 1,...,32 after the enrollment process.

4.2.2 Verification

After individual GMMs are trained using the maximum likelihood criterion to estimate the probability density functions P($x_i|M$) of the client speakers, the probability P($X|M$) that a test utterance X = $x_1, ..., x_L$ is generated by the model M is used as the utterance score where L is the number of frames in an utterance. This probability for the entire utterance is estimated by the mean log-likelihood over the sequence of frames that make up the whole utterance as follows:

$$S(X) = logP(X|M) = \frac{1}{L}\sum_{i=1}^{L} logP(x_i|M) \tag{4.2}$$

This utterance score is then used to make a decision by comparing it against a threshold that has been fixed for a desired EER. An alternative approach would be to generate a universal background model (UBM) and employ the log-likelihood ratio test as a 2-class classification problem where the log-likelihood score of the claimed model is compared to the world model. Here, the threshold is assumed to be 1 so the utterances are verified if the log-likelihood scores of the claimed model are higher than of the world model.

4.3 Proposed Method

4.3.1 Feature Extraction and Enrollment

As described in detail in Section 3.2.1, MFCCs are extracted from both enrollment and verification set utterances for the proposed system during the feature extraction stage. However, utterances in the enrollment set are aligned with the previously trained phonetic HMMs to segment the phonemes where the utterances of the verification set are left unaligned.

After MFCC features are extracted and the utterances are aligned in the enrollment set, these utterances are segmented into phonemes to create speaker specific phoneme codebooks. These codebooks C_i consist of 12-dimensional phoneme vectors F_p which are formed by taking the mean of the second state frames of

phoneme p from different utterances of user i. Formation of the phoneme vectors is described in detail in Section 3.2.1.

Phoneme codebooks for all speakers in the database are formed from the enrollment set utterances for the proposed TI system. Mean vectors are then calculated for each phoneme cluster and the resulting vector C_{pi} is assigned as the centroid to represent phoneme p of speaker i. At the end, the enrollment process is complete yielding 20 x N (20 phonemes x N speakers) 12-D centroids representing each phoneme of every speaker.

4.3.2 Verification

In order to verify a test utterance, MFCC based feature vectors are extracted first from all frames of the utterance. Then, a difference vector D is calculated by finding the nearest centroid for each frame vector in an utterance X of length L from the codebook of the claimed speaker. The distance metric used is the Euclidean distance without any modifications as explained in the previous chapter since the phoneme vectors have only 12 dimensions. Calculation of the difference vector D is shown below in equation 4.3.

$$D(l) = \min_{p \in P} \sqrt{\sum_{j=1}^{12} \left(X_l(j) - C_{pi}(j)\right)^2} \quad for \quad \forall l \in L \tag{4.3}$$

Here, the values in vector D are frame level distances in an utterance X. As the next step, utterance level score is calculated by taking the trimmed L2-norm of the difference vector D. This is done by calculating the Euclidean norm after discarding the highest valued dimensions in the vector, so that the remaining dimensions are more robust to certain noise artifacts. We have used the same percentage (10%) of discarded dimensions, as in section 3.2.1.

4.4 Database

To test our proposed text-independent speaker verification system and compare it with the baseline GMM-based system, we used the TIDIGITS database. The database consists of uttered digit sequences and is originally constructed for speech recognition. The details of the database are explained in Section 3.3.

For the implementation of the text-independent verification systems, the database of 163 speakers is divided into two sets for enrollment and verification. Enrollment set constitutes 1630 (10 utterances x 163 speakers) 3-digit and 1630 (10 utterances x 163 speakers) 4-digit speech samples. Speaker codebooks for the proposed method or GMM-based speaker models are created by using the frames of these utterances. Moreover, the utterances are not repetitive in the verification set in order to have

a fair distribution of existing phonemes for modeling speakers' vocal characteristics. In other words, original utterances of TIDIGITS database are employed for the proposed and GMM-based text-independent speaker verification systems.

On the other hand, verification set consists of 815 (5 utterances x 163 speakers) 7-digit utterances. As in the text-dependent verification system, 5 utterances in the verification set of each speaker are used only once for genuine attacks whereas they are used multiple times in a random manner when claiming other speakers' identities. Since the systems should be independent of text, there is no need to segment and concatenate the existing utterances as it has been done for the text-dependent system for the verification set.

4.5 Results

The performance evaluation of both text-independent speaker verification systems in this work is done by considering the false acceptance (FAR) and false rejection (FRR) rates during the tests. FAR is calculated as the ratio of falsely verified impostor utterances to the total number of impostor tests and FRR is calculated as the ratio of falsely rejected genuine utterances to the total number of genuine tests. EER and HER indicate Equal Error rate (where FRR and FAR are made equal by changing the acceptance threshold) the and Half Error Rate (average of FRR and FAR, when EER cannot be obtained).

Separate tests are conducted according to the classification method (proposed and baseline GMM method) and whether the forgers were selected from the same group as the person being forged, or the whole population (same group - SG or all groups - AG). As one can expect, the former (SG) is a more challenging scenario. For comparison, previously described text-dependent system (for known password scenario where the forger utters the same sequence of digits as in the claimed password) is also implemented with identical enrollment and verification sets as employed for the text-independent systems. For the results of the text-dependent system to be used as a benchmark, PCA with a linear classifier is utilized for comparison tests.

The results for the GMM-based method with 0.61% EER are superior to our proposed method for the text-independent verification case with 5.79% EER for the SG scenario; however the proposed text-dependent system performs better with 0.38% EER for the password-known scenario in which the forger knows the claimed password. For the AG scenario, again the GMM-based method with 0.31% EER is superior to our proposed method for the text-independent verification case with 5.79% EER; however the text-dependent system performs even better with 0.22% EER for the conditions described above. We list the results for both cases in Table 4.1.

Scenario	Same Group (SG)	All Groups (AG)
TD-PCA	0.39	0.22
TI-GMM	0.62	0.31
TI-Proposed	5.79	3.56

Table 4.1: Equal error rates are given for different classification methods (TD-PCA, TI-GMM, TI-Proposed) and whether the forger was selected from the same group as the person being forged, or the whole population.

Utilization of time dependent information (e.g. sequence of uttered digits) can be considered as the main strength of text-dependent verification systems as well as the acoustic features independent of text. Uttering a different sequence of digits than the enrolled pass-phrase ideally results in rejection of a genuine speaker in text-dependent verification systems, thus the role of temporal information is significant. For the case where the forger knows the password of the claimed speaker, extracted features from the utterance are only compared with corresponding templates/references (e.g. concatenated phoneme vectors in our case) and the role of the time dependent information is mainly reduced. Still, depending on the features extracted from the utterance (e.g. delta-MFCC or delta-delta-MFCC) some temporal information is used during verification. For the proposed text-dependent verification system, only MFCC features are utilized to extract fixed-length feature vectors and temporal information is discarded considerably. In this case, verification decision is considered to be done solely by the acoustic nature of the uttered text regardless of the length of the utterance. This is why our results for password-known scenario of text-dependent speaker verification tests are comparable with the results of text-independent verification tests.

Further tests were done to see if the performance would improve, if we knew the group (men/women/boys/girls) of the forged person. Note that this information can be derived from a person's age and gender which are public information. For this case, separate classifiers were trained for each group, using as forgers other people from the same group. In other words, if a man was being forged, we used a classifier trained with only information coming from adult male subjects. The results in Table 4.2 show that the error rates for men and women are much lower that of children (boys and girls groups) for all cases.

4.6 Summary

In this chapter, we proposed a text-independent speaker verification system using phoneme codebooks. These codebooks are generated by aligning the enrollment utterances using phonetic HMMs and creating MFCC-based fixed-length feature vectors to represent each phoneme. For verification, we define a distance metric

Group	TD-PCA	TI-GMM	TI-Proposed
boys	0.65	0.82	8.80
girls	0.82	0.73	6.37
men	0.36	0.22	2.67
women	0.18	0.34	4.92
average	0.41	0.43	4.96

Table 4.2: Error rates are given separately for each group and different classification methods (TD-PCA, TI-GMM, TI-Proposed). For these results, the classifiers are trained separately for each group where the forgers belong to the group of the claimed user.

measuring the total distance of a test utterance to the codebook normalized by the length of the utterance. For benchmarking, a GMM-based text-independent verification system and the proposed text-dependent speaker verification in chapter 3 are implemented using the same enrollment and verification dataset. The results of the text-dependent system (0.22% EER for the AG scenario) is superior to of the GMM-based system (0.31% EER for the AG scenario) whereas the proposed text-independent system yields worst results (5.79% EER for the AG scenario). Through creating phoneme codebooks, we tried to extract discriminative speaker information at the phoneme level;however, the results are not satisfactory. Instead of employing deterministic models such as our approach, generation of discriminative probabilistic models can yield better results. By using probabilistic models, no information can be lost as it was in our approach where information from only certain frames are utilized where the rest are neglected.

CHAPTER 5

Creating Multi-biometric Templates Using Fingerprint and Voice

In this chapter, we introduce a new implementation of the multi-biometric template framework of Yanikoglu and Kholmatov [12], using fingerprint and voice modalities. In this framework, two biometric data are fused at the template level to create a combined, multi-biometric template. The gain obtained through the combination in the proposed scheme is three-fold: increase in template security and privacy, increase in system security and cancelability. The constructed multi-biometric templates hide the constituent biometric data which is an important problem since any improperly disclosed biometric data is subject to identity theft. Furthermore, it is much less likely for the combined biometric template to reveal sensitive information. The multi-biometric templates constructed using the proposed method are also *non-unique* identifiers of the person, preventing cross-matching databases and searching a database with latent fingerprints or other unique identifiers, raising privacy concerns [12]. The performance of the biometric system is also increased since multiple biometrics are used in verification.

In the realization of this framework which is presented in this thesis, we use fingerprint and voice modalities, such that the fingerprint data is hidden using a voice pass-phrase. Fingerprint and voice are two of the most practical and commonly accepted biometrics. Furthermore, changing the password in text-dependent voice verification systems provides a cancelable biometric template.

The implementation of the multi-biometric template scheme provided in this thesis thus improves on the previous implementation done using two fingerprints by adding cancelability, as well as providing a new implementation supporting the scheme.

5.1 Previous Work

In this section, we provide a brief summary of the previous work in the areas of fingerprint verification as it relates to our work, and template security and privacy work. Along with the previous chapters on voice modality, this section should provide enough background to explain our proposed model.

5.1.1 Fingerprint Modality

Fingerprints have long been used for person verification and identification due to their immutability and uniqueness. Minutiae-based verification approaches are the most common, compared to ridge-based and correlation-based techniques [17]. The performance of minutiae-based fingerprint verification systems heavily depend on the minutiae extraction process done before minutiae alignment. Minutiae extraction is done using image processing operations that take advantage of the rich information available in the ridge structure of a fingerprint. Minutiae alignment, on the other hand, has to be done efficiently and should handle the non-linear deformations present in fingerprints.

Jiang and Yau use local structure to find the correspondence between two minutiae sets. They tested their method on a private database of 188 users and achieved an EER under 1% [58]. Jain et al proposed an alignment based algorithm were the ridge information is employed to align the minutiae sets and a bounding box was proposed to match aligned minutiae [3]. Further improvements for this method have been proposed by He et al where the EER is decreased from around 3-4% to 1-2% in a database of 100 users [59]. Tico and Kuosmanen employed orientation field information of the fingerprint pattern to create a fingerprint representation scheme [34] and Ratha et al proposed a matching technique based on the graph representations of the query and template fingerprints, constructed using their respective minutiae features [39].

5.1.2 Template Security and Privacy

Template security and privacy are main concerns in building biometric systems for numerous reasons. Protecting the biometric information against disclosure of the biometric data itself, or other sensitive information, as well as preventing cross-matching of databases is an active research area. Numerous architectures have been proposed in recent years, aiming to protect biometric templates stored in central repositories [20, 2, 37, 25]. Among those, *fuzzy vault* technique is one of the most widely used method where points obtained from a biometric modality (e.g. fingerprint) are stored with randomly generated chaff points [2]. A user has to provide a certain number of minutiae points to unlock the vault created by the reference fingerprints minutiae set given during the enrollment session. The scheme is based on the difficulty of the polynomial reconstruction problem. There have been many implementation of the Fuzzy Vault scheme using different biometric modalities such as fingerprint [53, 50, 7], signature [4], as well as work showing the weaknesses of the Fuzzy Vault scheme [6, 56].

Yanikoglu and Kholmatov proposed another method based on the idea of com-

bining multiple biometrics in order to increase both privacy and security [12]. In the given implementation of the general idea, minutiae points from two distinct fingers of the same person were combined to create a *multi-biometric template* which is later shown to be more unique, hence more privacy preserving. They also showed that the system provides higher level of security as well, as it verifies both fingerprints. In this chapter, an implementation of the multi-biometric template framework with fingerprint and voice modalities is presented.

5.1.3 System Security

System security is another main concern in designing biometric systems beside privacy protection. High security applications require very low error rates and uni-modal biometric systems are not always satisfying in that regard, while multi-modal biometric systems have proven to be useful in increasing the system security.

In literature, the combination of multiple biometrics mostly take place at the matching score or decision level [47, 17]. Some examples of research in multi-biometric systems are the following: Brunelli and Falavigna use the hyperbolic tangent for normalization and weighted geometric average for fusion of voice and face biometrics [44]. These modalities have also been fused by Ben-Yacoub et al by considering several strategies such as support vector machines, tree classifiers and multilayer perceptrons [49]. Kittler et al have experimented with fusion techniques for face and voice on the matching score level [22]. Hong and Jain proposed an identification system using face and fingerprint where the database is pruned via face matching before fingerprint matching [28].

5.2 Proposed Method

For this work, we implemented the multi-biometric template framework [12] using fingerprint and voice-modalities. Voice and fingerprint data of individuals are fused at the template level by combining minutiae points obtained from fingerprints with artificially constructed “minutiae“ points obtained from the utterance, as described in the following subsections. We show that the system security is higher compared to single biometric counterparts using only fingerprint or only voice.

5.2.1 Feature Extraction from Fingerprint

For fingerprints, minutiae points from the ridge endings and bifurcations on the fingerprint pattern are used as features in our work. In literature, there are several methods proposed for automatic minutiae extraction [1, 36], which commonly follow well-known image enhancement, binarization and thinning steps. Automatic

detection of minutiae points can sometimes result in missed or spurious minutiae. Thus, minutiae points found through image processing operations are later verified using various post-processing techniques [33] in some systems. After minutiae extraction, finger print verification involves minutiae alignment and matching. In that process, the challenges are caused by non-linear deformations of the fingerprint, as well as missing and spurious minutiae.

Since the aim of this work is to build a multimodal biometric system with information fusion at the template level, we preferred to utilize fingerprint images with manually labeled minutiae points. This is done in order to reduce errors which might arise from the minutiae extraction process. For template generation, the minutiae points from users (roughly 30 minutiae points on average) are stored in a 2-dimensional plane with their x and y coordinates as features. This process in illustrated in Figure 5.1.

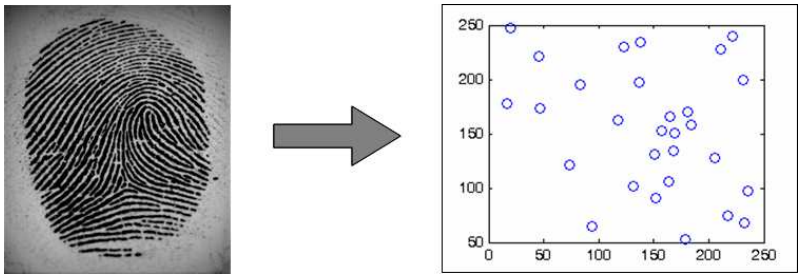

Figure 5.1: The minutiae points from fingerprints are extracted manually and stored in a 2 dimensional plane with their x and y coordinates as features.

5.2.2 Feature Extraction from Voice

As described in previous chapters, short spectra of speech signals give information about both the spoken words and the voice of the speaker. Therefore, we use the Mel Frequency Cepstral Coefficients (MFCCs) as the voice features. The feature extraction process explained in 3.2.1 is summarized below for convenience.

All utterances (reference or query) are first aligned using an HMM of the spoken password of the claimed identity. These pass-phrase models are formed by concatenating the HMMs of its constituent phonemes. They are speaker-independent models, trained using a separate part of the database. After the alignment, frames which correspond only to the middle state of each phoneme are kept, while the remaining frames (those corresponding to the 1^{st} and 3^{rd} states) are deleted. This is done to use only steady-state portions of each phone, and eliminate the start and end states that are more affected by the neighboring phones. We then calculate the mean feature vector of cepstral coefficients for each phoneme. After the

calculations, each phoneme p is represented by a 12-dimensional mean vector F_p. Using this method, fixed-length feature vectors are extracted from varying-length utterances consisting of the same digits. Extraction of voice features is illustrated in Figure 5.2. The concatenation of the mean vectors of the phonemes then forms the feature vector for the entire utterance.

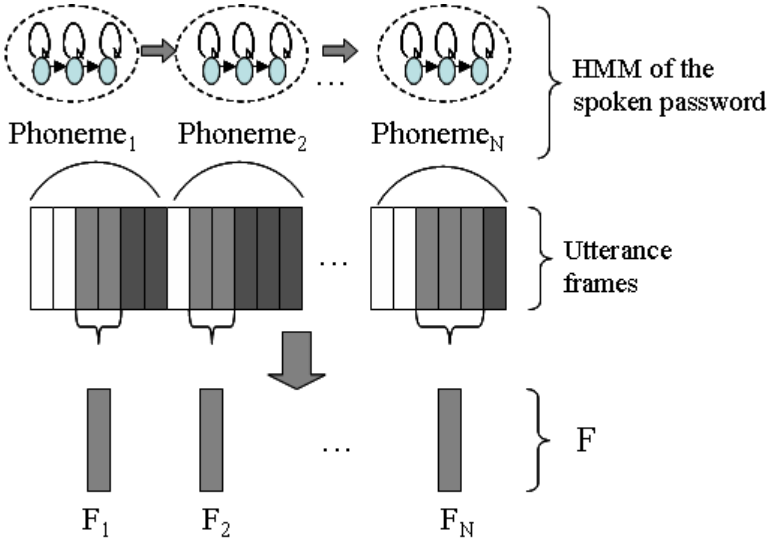

Figure 5.2: Alignment with 3-stage HMMs: Previously trained 3-stage phonetic HMMs are used in aligning an utterance, to find the correspondence between individual frames and phonemes. Phoneme 1-N indicate the phonemes that occur in spoken password. Levels of gray (white-gray-dark gray) indicate the 3-stages within a phoneme.

5.2.3 Multi-biometric Template Generation

The motivation behind combining fingerprint and voice is to hide fingerprint data along with a voice pass-phrase and store the combined multi-biometric template. In order to hide the fingerprint data along with a voice pass-phrase, both features need to be combined in the same feature space. Since fingerprint features are already in the Euclidean space, we transform the voice features into the same space to create the multi-biometric templates. First, 12-dimensional mean vectors of each phoneme are concatenated to form a feature vector F to represent the entire pass-phrase after the voice feature extraction process. Therefore, the F vector of an utterance with N phonemes is $N \times 12$ dimensional. Next, the F vector representing an utterance is binarized using a global threshold t of -3.5 for females and -1 for males. The binarization process is done by assigning bits to each dimension of the F vector according to its value (if the MFC coefficient is above the threshold, 1 is assigned to that dimension). This way, every utterance is now represented with a

bit string B with length $N \times 12$. The t values are determined in order to have equal number of 1 and 0 bits for all speakers. Binarization process is illustrated in the top two rows of Figure 5.3.

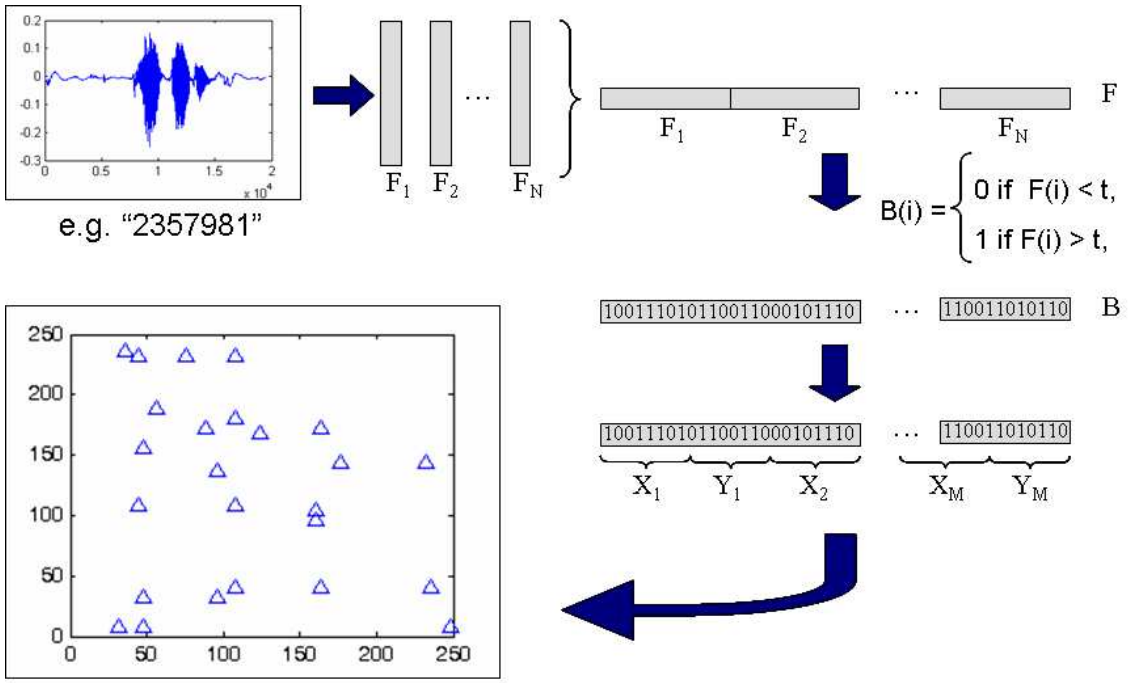

Figure 5.3: Minutiae point generation from voice: mean feature vectors from the previously aligned utterances are concatenated and binarized according to a predetermined threshold, then the bit string is divided into chunks of 8 bits to obtain the artificial utterance points (X_i, Y_i).

For using voice as a biometric, it is necessary to go through a speaker specific enrollment session. In this work, there is only one enrollment session where 10 utterances of the chosen password are collected from each user. The binary feature vectors B_i, generated from the reference utterances of a user, are combined so as to obtain a single reference feature vector, by majority voting of the bits in every dimension. This single feature vector represents the voice template of a single speaker, dependent on the chosen password. This single reference feature vector is also referred as the **binary feature descriptor** of the user.

Finally, to map the resulting binary voice template onto the 2-dimensional space (the Euclidean space) of minutiae points, we followed a few different methods with roughly similar results: in one, we divided each binary feature descriptor of the users into groups of 16 and used each 16 bits to generate one point (X,Y) in the Euclidean space. Of these 16 bits, 2 decimal numbers with values ranging from 0 to 255 are extracted from 8 bits each and mapped to a 2-dimensional point. The reason for using 8 bits to represent a dimension in the Euclidean space is that all the fingerprint minutiae data fall in a square frame of side-length 256. This way,

when the artificially generated points from voice and fingerprint minutiae are fused together, they cover the same region in the 2-dimensional Euclidean space.

As a second method, groups of 12 bits from the binary feature descriptors of users are employed to create more points in the 2-D Euclidean space instead of 16. Originally, if points are created from 12 bits, they fall in a square frame with side-length 64. This time, the points are scaled to fall inside a square frame with a side-length of 256 to cover the same region with the fingerprint minutiae. The process of creating artificial minutiae points is shown in the bottom row of Figure 5.3 above. The points created by grouping the bits of the binary descriptors of the combined reference utterance or test utterances will be referred as the "voice minutiae" of users from now on.

For this work, 6-digit pass-phrases are used. Since each digit in the English language is composed of roughly 3 phonemes, 18 phonemes are present on average in each pass-phrase. This means, roughly 18 12-dimensional feature vectors are extracted per utterance of a password. Thus, binary feature descriptors are composed of $18 \times 12 = 216$ bits, yielding roughly 14 voice minutiae points per pass-phrase if they are grouped by 16 bits (8 bits for 1 point). To complete the enrollment or template generation phase, minutiae points extracted from the fingerprint (A) are fused with the reference voice minutiae (B) of the user to form the multi-biometric template (A+B) for the user, **without** any labels indicating the origin of the points. Fusion of biometric data is illustrated in Figure 5.4.

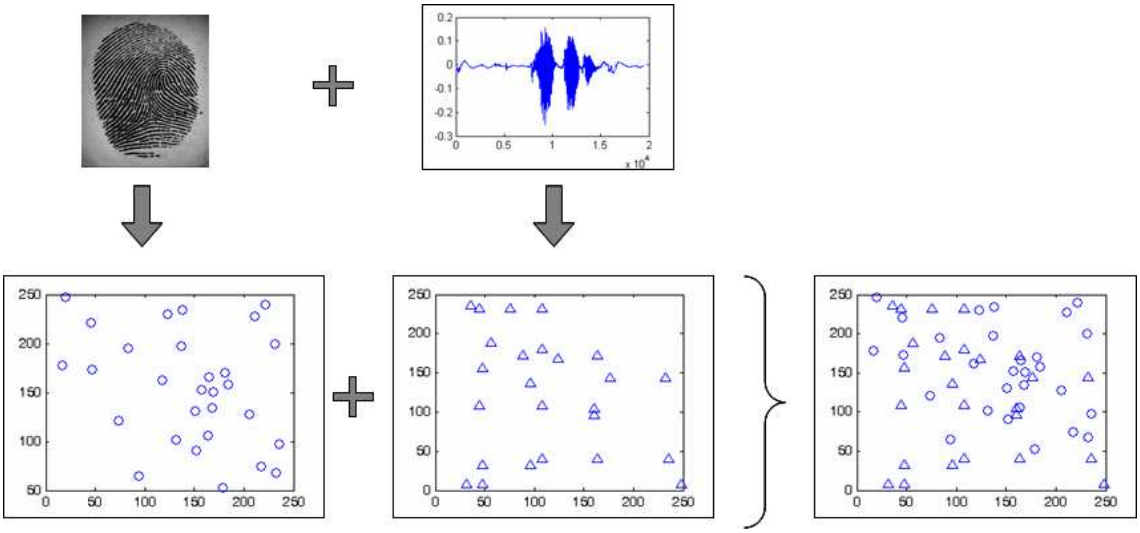

Figure 5.4: Template level fusion of biometric data: Minutiae points from the fingerprint and artificial points generated from voice are combined together in a user template. The points are marked as to indicate the source biometric, but this information is not stored in the database.

5.2.4 Verification

When a user comes for authentication, he or she is authenticated using both the fingerprint and voice modalities, in order. First, the fingerprint minutiae (A') are extracted and matched against the template (A+B) of the claimed user, consisting of both fingerprint and voice minutiae. The automatic matching is done via a simple matching algorithm finding the best alignment over all translations and rotations, allowing for some elastic deformation of the fingerprint (accepting two points as matching if they are within a small threshold in this alignment). After the alignment, the matched points are deleted from the template, leaving non-matched points in the template, resulting in (A+B-A'). Note that ideally these are only the points generated from the utterance of the claimed user. Matching of the test fingerprint is illustrated in Figure 5.5.

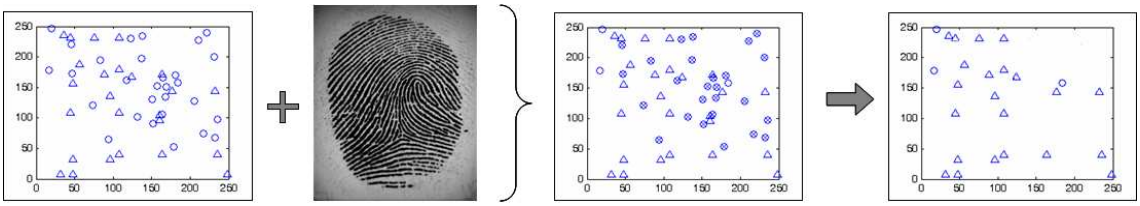

Figure 5.5: Illustration of the first phase of the verification, where the test fingerprint is matched with the user template shown on the left. Matched points of the template are marked with a cross and removed in the rightmost part of the figure.

Next, the test user provides the pass-phrase of the claimed user. If the pass-phrase is unknown, it is guessed. The *F'* vector representing the test utterance is binarized using a global threshold t and the binary feature descriptor is divided into chunks of 16 or 12 bits to be mapped to the Euclidean space, as described in the previous sections, resulting in the second individual template (B').

Notice that even a genuine utterance may have a small number of bits different from that of the binary feature descriptor used in the stored template. Hence, the matching of the voice minutiae with the remaining points in the template (A+B-A') cannot take place in Euclidean space because small variations may result in unacceptable distances, depending on the position of the erroneous bit based on which the voice minutiae was generated. If there is an erroneous bit in the most significant bit of the x or y coordinate, that voice minutia will be far away from the reference voice minutia in the Euclidean space. Thus, we carry the matching of the test voice minutiae (B') with the remaining points in the template (A+B-A'), in the Hamming space. This way, a small number of bit errors caused by the variations in the user's voice feature will cost only a few (probably 1 or 2) bits in Hamming space. To do this, all the remaining points in the template (A+B-A') are

mapped back to the Hamming space by inverting the mapping of the binary feature descriptor onto the Euclidian space. In other words, we concatenate the x and y coordinates of a voice minutiae point, after doing a decimal-to-binary conversion, to obtain the 12 or 16 dimensional original feature from which the voice minutiae was extracted. To be considered a match, the Hamming distance between two bit strings should be less than or equal to a threshold (2-bit errors are accepted in this work for 16-bit strings and 1-bit errors are accepted for 12-bit strings). Then, matched voice minutiae points are deleted from the template (A+B-A'-B') of the claimed identity. Matching of the test voice minutiae is illustrated in Figure 5.6.

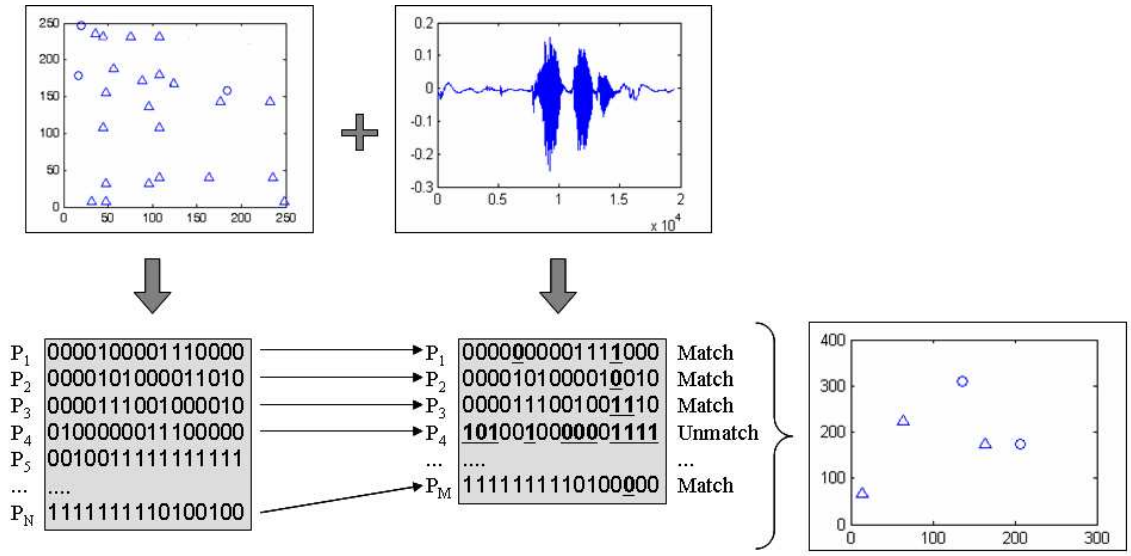

Figure 5.6: Illustration of the second phase of the verification where the utterance is matched with the remaining points in the user's template. Matched points are removed, showing here a successful match.

5.2.5 Matching Decision

We have used two score metrics for the verification decision in this work. One is the same with the work of Yanikoglu and Kholmatov [12] and the second is a slight variation of it.

For the first metric, which is the Jaccard index, we consider the overlap between the remaining points and the points in the second template (voice): the person is authenticated if a high percentage of the points involved in the second phase of the verification (remaining points from the template and the minutiae points of the voice) is matched.

$$score = Jaccard\left(\left(A + B - A'\right), B'\right) = 2 \times \frac{\left|\left(A + B - A'\right) \cap B'\right|}{\left|\left(A + B - A'\right) + B'\right|} \qquad (5.1)$$

In case A' matches A perfectly and B' matches B perfectly for this metric, the resulting score with this metric is 1, showing a good match. If A' was not successfully matched, it would be reflected in the final score since many minutiae points would be left unmatched, making the denominator large. If B' was not successfully matched, the numerator would be small.

As an alternative, we came up with a new score metric derived from the first one described above. The key here is that since the number of voice minutiae to be extracted from the test utterance is known, this information can be used to **roughly** measure the independent performance of the voice matching process to increase the overall success of the score metric. Note that we cannot measure the independent performance of the voice matching precisely because we do not know exactly which points on the Euclidean space are generated artificially from voice –just their count. So the following derived score metric takes into account the fingerprint and voice minutiae counts (A' and B' respectvely):

$$score' = \frac{|(A+B) \cap A'|}{|A|} + \frac{|(A+B-A') \cap B'|}{|B|} \tag{5.2}$$

Thus, we basically combine the individual performances of fingerprint and voice matching processes. If one of these matching processes perform poorly, the overall score decreases linearly due to the addition of matching performances.

5.3 Database

We employed the TIDIGITS database as the utterance database for this work. The database consists of uttered digit sequences and is originally constructed for speech recognition. The details of the database are further explained in Section 3.3. Utterances of 100 (50 men, 50 women) speakers who are randomly chosen among 113 (56 men, 57 women) speakers are processed for multi-biometric template creation and matching.

In order to obtain multiple sequences of the same password, as needed in a text-dependent speaker verification, we segmented the utterances of 100 speakers in the database into digits, using the previously described HMMs. This resulted in 16 or more utterances of the same digit by each user. Ten of each of these digits are used for enrollment (reference feature vectors extraction followed by reference binary feature descriptor generation) and 5 for verification (test feature vectors extraction followed by test binary feature descriptor generation for genuine and forgery verification tests).

To create the enrollment set, we randomly picked a 6-digit password for each user in the database and created artificial utterances of this password by combining segments of the constituent digits. After creating the enrollment set, genuine test

cases were created using the unused digits of the same user (5/16+). For this chapter, only password-known tests are conducted to simulate the scenario of a stolen password. Therefore, forgery utterances are created using the **same** digit sequence as the claimed password for the tests.

The enrollment set thus consists of a total of 1000 (10 utterances x 100 speakers) reference passwords where each recorded digit is used only once. The genuine test set contains 500 (5 utterances x 100 speakers) genuine passwords constructed from genuine, segmented digit recordings. The forgery test set contains 4900 (50x49x2 for men and women) forgery passwords constructed from segmented digit recordings of other people. Here, each speaker forges every other speaker in the same gender with only one utterance of the claimed password. In other words, each speaker is forged by all the remaining speakers of the same gender who knows his/her password.

The fingerprint database on the other hand consists of 200 fingerprints from 100 individuals (2 imprints per person). We matched each person with a randomly chosen speaker from the utterance database, to link these two unrelated databases and create a multi-biometric database. We then created 100 multi-biometric templates using one fingerprint of a person from the fingerprint database and the binary feature descriptor obtained from the 10 reference feature vectors of the matching person from the utterance database following the feature extraction processes. In reference to the template, we will refer to these two unrelated people whose fingerprint and voice data are fused as the "user" from now on.

5.4 Results

The performance evaluation of the biometric verification system proposed in this work is done by considering the false acceptance and false rejection rates of the system. The verification tests are done using 3 different scenarios: In the first scenario (FF), we tested the case where both the test fingerprint and the test utterance is a forgery. Hence, we had 5 genuine tests (1 fingerprint x 5 utterances) and 49 forgeries (1 fingerprint x 49 utterances) for each user, generated by false fingerprints and forgery utterances.

For the second scenario (FG), we tested the case where the fingerprint is a forgery, but the utterance is genuine. This may be the situation where the forger may have recorded a sample of the user whose voice is stored in the multi-biometric template. Hence, the EER is calculated by comparing the results from 5 genuine tests (1 fingerprint x 5 utterances) with 49 (49 fingerprints x 1 utterance randomly chosen from the genuine utterances of the claimed user) forgery tests for each user, generated by false fingerprint but genuine voice samples.

For the last scenario (GF), we tested the case where the fingerprint is genuine, but the utterance is forgery. This may be the situation where the forger may have acquired a silicon fingerprint of the user whose biometric is stored in the multi-biometric template. Hence, the EER is calculated by comparing the results from 5 genuine tests (1 fingerprint x 5 utterances) with 49 (1 fingerprint x 49 utterances) forgery tests for each user, generated by genuine fingerprint but false utterances.

In addition to different test scenarios, further tests are conducted with different bit string lengths for segmenting the binary feature descriptor generated from the voice feature vectors of users. Moreover, all these tests are repeated for 2 different score metrics. The results for this work can be seen in Table 5.1, 5.2 and 5.3 for male users and female users separately for all 3 test scenarios and score metrics.

Gender	Original Score Metric		Proposed Score Metric	
	12-bit	16-bit	12-bit	16-bit
Male (50)	4.50%	5.15%	3.90%	0.67%
Female (50)	3.05%	4.08%	1.32%	0.87%
Average (100)	3.78%	4.62%	2.61%	0.77%

Table 5.1: Scenario 1 (FF) - The results are given for the case where both the test fingerprint and the test utterance is a forgery for the impostor attempts.

Gender	Original Score Metric		Proposed Score Metric	
	12-bit	16-bit	12-bit	16-bit
Male (50)	23.50%	24.20%	10.40%	6.30%
Female (50)	17.20%	21.60%	3.90%	5.90%
Average (100)	20.35%	22.90%	7.15%	6.10%

Table 5.2: Scenario 2 (FG) - The results are given for the case where the fingerprint is a forgery, but the utterance is genuine is a forgery for the impostor attempts.

Gender	Original Score Metric		Proposed Score Metric	
	12-bit	16-bit	12-bit	16-bit
Male (50)	10.60%	8.70%	17.20%	12.70%
Female (50)	9.90%	9.80%	15.10%	12.80%
Average (100)	10.25%	9.25%	16.25%	12.75%

Table 5.3: Scenario 3 (GF) - The results are given for the case where the fingerprint is genuine, but the utterance is forgery.

As can be see in Table 1, in a forgery attempt with a false fingerprint and false utterance, the equal error rate is only 0.77% for the new score metric. This rate increases to 6.1% when the attacker has access to the utterance of the user. Notice that the case when the attacker has access to the fingerprint of the user, the error rate significantly increases; however, even then about 87% of the attacks are

repelled. For this case, the EER for the original score metric is 25% lower with respect to the EER for the proposed score metric (12.75%).

For comparison, in our work for unimodal speaker verification system as described in Chaper 3, average EER for men and women turns out to be 0,34% with similar voice feature extraction algorithm for the same speaker database. The average results obtained for impostors with mismatched fingerprint and voice (0.77%) for this work are not lower than the unimodal system counter to what was expected. Employment of trained classifiers with multiple reference vectors in the feature extraction stage for the proposed unimodal text-dependent verification system yields better results than the proposed system using multi-biometric templates.

Furthermore, we have experimented with different quantization level and different partitioning of the data, as well as using PCA for projecting the 6-dimensional parts of the feature vector into 2-dimension (x,y) with similar initial results. We are experimenting with them further to obtain more reliable 2-dimensional features from voice. For future work, we also plan to measure how the multi-biometric template scheme preserves privacy, by calculating recall and precision rates for retrieving templates given only one of the biometric modalities, as previously done by Yanikoglu and Kholmatov [12].

5.5 Summary and Contributions

In this chapter, we introduced a new implementation of the multi-biometric template framework of Yanikoglu and Kholmatov [12], using fingerprint and voice modalities. In this framework, two biometric data are fused at the template level to create a combined, multi-biometric template, in order to increase both security and privacy of the system, at the same time.

In addition to the first implementation of this framework, which used two fingerprints and showed increases in both security and privacy, the implementation presented here also provides cancelability. Cancelability of the multi-biometric template is achieved by changing the pass-phrase uttered by the speaker, since the generated voice minutiae depends on the pass-phrase comprised of a unique sequence of phonemes.

Our work for unimodal speaker verification system as described in Chapter 3 yields slightly better results (0.34% EER) in comparison with the multi-biometric method introduced here (0.77% EER), for the same speaker database. However, the results are close and the gain in privacy may be preferred to some loss in equal error rate of the resulting system. Employment of trained classifiers with multiple reference vectors in the feature extraction stage for the proposed unimodal text-dependent verification system can be listed as reasons for this outcome.

CHAPTER 6

Contributions and Future Work

In this thesis, we present three new biometric verification systems based mainly on voice modality. First, we propose a text-dependent (TD) system where acoustic features are extracted from individual frames of the utterances, after they are aligned via speaker-independent, phonetic HMMs. Second, a text-independent (TI) speaker verification method is implemented, inspired by the feature extraction method utilized for our text-dependent system. Third, we introduce a new implementation of the multi-biometric template framework of Yanikoglu and Kholmatov [12], using fingerprint and voice modalities.

In the TD system presented in Chapter 3, we present a new method for text-dependent speaker verification through extraction of fixed-length feature vectors from utterances. The system is faster and uses less memory as compared to the conventional HMM-based approach, while having state-of-the-art results. In our system, we only use a single set of speaker-independent monophone HMM models. This set is used for alignment, whereas for the conventional HMM-based approach, an adapted HMM set for each speaker is constructed in addition to a speaker independent HMM set (also called universal background model in that context). This requires much higher amount of memory as compared to the proposed approach. In addition, during testing only a single HMM alignment is required as compared to two HMM alignments using a universal background model and a speaker model for the conventional approach. Thus, verification is also faster with the approach introduced in this thesis. The results from text-dependent verification system (0.61 and 0.39% EER) may be compared to the results of Ramasubramanian et al. (under 0.1% EER) who have used the same database under similar conditions [54]. They use multiple utterances of the same digit to create digit templates which are used in verifying utterances of a known digit sequence. However, their EER is lower through cohort normalization using a closed-set verification scenario. In other words, the decision mechanism does not only know about the similarity of the query utterance, but also the similarity of the forgery utterances, which significantly improves verification performance. For future studies, delta coefficients of acoustic features can

be employed to include duration information of utterances. Besides, several other features such as LPC and their combinations can also be used to decrease the EER of the text-dependent system. Similarly, the results by Subramanya et al [10] who created a database suitable for text-dependent verification from the original YOHO database may also be compared to ours. However, their results of 0.26% should be compared to the average of "men" and "women" groups (0.34%) in our work since "boys" and "girls" groups do not exist in the YOHO database.

In Chapter 4, we propose a text-independent speaker verification system using phoneme codebooks. These codebooks are generated by aligning the enrollment utterances using phonetic HMMs and creating MFCC-based fixed-length feature vectors to represent each phoneme. For verification, we define a distance metric measuring the total distance of a test utterance to the codebook normalized by the length of the utterance. For benchmarking, a GMM-based text-independent verification system and the proposed text-dependent speaker verification in chapter 3 are implemented using the same enrollment and verification dataset. The results of the text-dependent system (0.22% EER for the AG scenario) is superior to that of the GMM-based system (0.31% EER for the AG scenario), whereas the proposed text-independent system yields worst results (5.79% EER for the AG scenario). Through creating phoneme codebooks, we tried to extract discriminative speaker information at the phoneme level; however, the results are not very satisfactory. Instead of employing deterministic models such as our approach, generation of discriminative probabilistic models may yield better results. An alternative text-independent system in the spirit of this work can be implemented by using a generic phone recognizing HMM as a future direction. After the phone recognizer obtains an alignment of frames to phonemes, a fixed-length feature vector representing the utterance can be formed. Then as in the text-dependent method, simple distance-based algorithms can be used for person recognition.

In Chapter 5, we introduce a new implementation of the multi-biometric template framework of Yanikoglu and Kholmatov [12], using fingerprint and voice modalities. In this framework, two biometric data are fused at the template level to create a combined, multi-biometric template, in order to increase both security and privacy of the system at the same time. In addition to the first implementation of this framework, which used two fingerprints and showed increases in both security and privacy, the implementation presented here also provides cancelability. Cancelability of the multi-biometric template is achieved by changing the pass-phrase uttered by the speaker, since the generated voice minutiae depends on the pass-phrase comprised of a unique sequence of phonemes. Our work for unimodal speaker verification system as described in Chapter 3 yields slightly better results (0.34% EER) in comparison with the multi-biometric method introduced here (0.77% EER), for

the same speaker database. However, the results are close and the gain in privacy may be preferred to some loss in equal error rate of the resulting system. Employment of trained classifiers with multiple reference vectors in the feature extraction stage for the proposed unimodal text-dependent verification system can be listed as reasons for this outcome. Furthermore, we have experimented with different mapping techniques, including using PCA for projecting the MFCC coefficients into 2-dimension (x,y) with similar but slightly worse results. We are experimenting with them further to obtain more reliable 2-dimensional features from voice with high inter-class and low intra-class variation. Through more efficient mapping methods, larger amount data from MFCC coefficients can be saved and constructing more discriminative multi-biometric templates can be possible. For future work, we also plan to measure how the multi-biometric template scheme preserves privacy, by calculating recall and precision rates for retrieving templates given only one of the biometric modalities, as previously done by Yanikoglu and Kholmatov [12].

REFERENCES

[1] S. Pankanti R. Bolle A. Jain, L. Hong. An identity authentication system using fingerprints. In Proc. of the IEEE, volume 85, pages 1365–1388, Sep. 1997.

[2] M. Sudan A. Juels. A fuzzy vault scheme. In Proc. of IEEE Int. Symp. on Information Theory, volume 408, 2002.

[3] R. Bolle A. K. Jain, L. Hong. On-line fingerprint verification. IEEE Trans. on Pattern Analysis and Machine Intelligence, 19(4):302–314, 1997.

[4] B. A. Yanikoglu A. Kholmatov. Biometric cryptosystem using online signatures. In Int. Symp. on Computer and Information Sciences, volume 4263, pages 981–990, Nov. 2006.

[5] B. Yanikoglu A. Kholmatov. Identity authentication using improved online signature verification method. Pattern Recognition Letters, 26(15):2400–2408, Nov. 2005.

[6] B. Yanikoglu A. Kholmatov. Realization of correlation attack against fuzzy vault. In Security, Forensics, Steganography, and Watermarking of Multimedia Contents X, Electronic Imaging, Jan. 2008.

[7] E. Savas A. Levi A. Kholmatov, B. A. Yanikoglu. Secret sharing using biometric traits. In SPIE Biometric Technology For Human Identification III, volume 6202, Apr. 2006.

[8] D. B. Rubin A. P. Dempster, N. M. Laird. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, 39(1):1–38, Feb. 1977.

[9] A. K. Jain A. Ross. Mutlimodal biometrics: an overview. In Proceeding of European Signal Processing Converence, pages 1221–1224, Sep. 2004.

[10] A. C. Surendran P. Nguyen M. Narasimhan A. Acero A. Subramanya, Z. Zhang. A generative discriminative framework using ensemble methods for text-dependent speaker verification. In IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pages 225–228, 2007.

[11] N. Scheffer J. F. Bonastre J. S. D. Mason B. G. B. Fauve, D. Matrouf. State-of-the-art performance in text-independent speaker verification through open-source software. IEEE Trans. on Speech and Audio Processing, 15(7):1960–1968, Sep. 2007.

[12] A. Kholmatov B. Yanikoglu. Combining multiple biometrics to protect privacy. In Proceedings of ICPR-BCTP Workshop, 2004.

[13] J.M. Zachariah C.S. Gupta B. Yegnanarayana, S.R.M. Prasanna. Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system. IEEE Trans. on Speech and Audio Processing, 13(4):575–582, Jul. 2005.

[14] D. Yuk C. Che, Q. Lin. An hmm approach to text-promted speaker verification. In IEEE Int. Conf. on Acoustics, Speech and Signal Processing, volume 2, pages 673–676, 1996.

[15] R. B. Dunn D. A. Reynolds, T. F. Quatieri. Speaker verification using adapted mixture models. Digital Signal Processing, 10:181–202, 2000.

[16] G. Stockman D. Colbry, F. Oki. 3d face identification - experiments towards a large gallery. In SPIE Conf. on Defense and Security, 2008.

[17] A. K. Jain S. Prabhakar D. Maltoni, D. Maio. Handbook of Fingerprint Recognition. Springer, New York, 2003.

[18] Q. Li F. Monrose, M. K. Reiter and S. Wetzel. Cryptographic key generation from voice. In IEEE Symposium on Security and Privacy, May 2001.

[19] G. D. Forney. The viterbi algorithm. In Proc. of IEEE, volume 61, pages 268–278, 1973.

[20] B. Matt G. Davida, Y. Frankel. On enabling secure applications through on-line biometric identification. In Proc. of the IEEE Symp. on Security and Privacy, pages 148–157, 1998.

[21] W. B. Mikhael G. Zhou. Speaker identification based on adaptive discriminative vector quantisation. IEE Proc. on Vis. Image Signal Processing, 153(6):754–760, Dec. 2006.

[22] R.P.W. Duin J. Matas J. Kittler, M. Hatef. On combining classifiers. IEEE Trans. on Pattern Analysis and Machine Intelligence, 20(3):226– 239, Mar. 1998.

[23] M. Neeracher K. E. A. Silverman J. R. Bellegarda, D. Naik. Language-independent, short-enrollment voice verification over a far-field microphone. In IEEE Int. Conf. on Acoustics, Speech and Signal Processing, volume 1, pages 445–448, May 2001.

[24] I. T. Jollife. Principal Component Analysis. NY: Springer Verlag, 1986.

[25] P.Tuyls J.P. Linnartz. New shielding functions to enhance privacy and prevent misuse of biometric templates. In Proc. of the 4^{th} Int. Conf. on Audio and Video Based Biometric Person Authentication, pages 393–402, 2003.

[26] G. R. Sell L. E. Baum. Growth functions for transformations on manifolds. Pacific Journal of Mathematics, 27(2):211–227, 1968.

[27] G. Soules N. Weiss L. E. Baum, T. Petrie. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. Annals Institute of Statistical Mathematics, 41(1):164–171, Apr 1970.

[28] A.K. Jain L. Hong. Integrating faces and fingerprints for personal identification. IEEE Trans. on Pattern Analysis and Machine Intelligence, 20(2):1295–1307, Dec. 1998.

[29] J. A. Egon L.E. Baum. An inequality with applications to statistical estimation for probabilistic functions of a markov process and to a model for ecology. Bulletin of American Meteorological Society, 73:360–363, 1967.

[30] Z. Yao B. Dai M. Liu, Y. Xie. A new hybrid gmm/svm for speaker verification. In IEEE 18^{th} Int. Conf. on Pattern Recognition, volume 4, pages 314–317, 2006.

[31] B. Mak M. Mak, R. Hsiao. A comparison of various adaptation methods for speaker verification with limited enrollment data. In IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pages 929–933, 2006.

[32] J. Epps M. Nosratighods, E. Ambikairajah. Speaker verification using a novel set of dynamic features. In IEEE 18^{th} Int. Conf. on Pattern Recognition, volume 4, pages 266–269, 2006.

[33] P. Kuosmanen M. Tico. An algorithm for fingerprint image post-processing. In Proc. of the 34^{th} Asilomar Conference on Signals, Systems and Computers, volume 2, pages 1735–1739, 2000.

[34] P. Kuosmanen M. Tico. Fingerprint matching using an orientation-based minutia descriptor. IEEE Trans. on Pattern Analysis and Machine Intelligence, 25(8):1009–1014, Aug. 2003.

[35] J. H. Connell R. M. Bolle N. K. Ratha, S. Chikkerur. Generating cancelable fingerprint templates. IEEE Trans. on Pattern Analysis and Machine Intelligence, 29(4):561–572, Apr 2007.

[36] A. Jain N. Ratha, S. Chen. Adaptive flow orientation based feature extraction in fingerprint images. Pattern Recognition, 28:1657–1672, 1995.

[37] R. Bolle N. Ratha, J. Connell. Enhancing security and privacy in biometrics-based authentication systems. IBM Systems Journal, 40:614–634, 2001.

[38] P. C. Ching N. Zheng, T. Lee. Integration of complementary acoustic features for speaker recognition. IEEE Signal Processing Letters, 14(3):181–184, Mar. 2007.

[39] V.D. Pandit N.K. Ratha. Robust fingerprint authentication using local structural similarity. In 5^{th} IEEE workshop on Applications of Computer Vision, pages 29–34, 2000.

[40] J. H. L. Hansen P. Angkititrakul. Discriminative in-set out-of-set speaker recognition. IEEE Trans. on Speech and Audio Processing, 15(2):498–508, Feb. 2007.

[41] A. K. Nandi P. Day. Robost text-independent speaker verification using genetic programming. IEEE Trans. on Audio, Speech and Language Processing, 15-1(1):285–295, Oct. 2007.

[42] C.H.Lee Q.Zhou F.K. Soong Q. Li, B.H. Juang. On speaker authentication. In IEEE Workshop on Automatic Identification Advanced Technologies, pages 3–6, Nov. 1997.

[43] S. Kwong Q. Y. Hong. A discriminative training approach for text-independent speaker recognition. Signal Processing, 85:1449–1463, 2005.

[44] D. Falavigna R. Brunelli. Person identification using multiple cues. IEEE Trans. on Pattern Analysis and Machine Intelligence, 17(10):955–966, Oct. 1995.

[45] D. G. Stork R. O. Duda, P. E. Hart. Pattern Classification 2^{nd} Ed. Wiley, 2000.

[46] R. Ramachandran R. J. Mammone R. P. Ramachandran, K. R. Farrel. Speaker recognition - general classifier approaches and data fusion methods. Pattern Recognition, 35:2801–2821, 2002.

[47] A. Mink M. Indovina A. Jain R. Snelick, U. Uludag. Large-scale evaluation of multimodal biometric authetication using state-of-the-art systems. IEEE Trans. on Pattern Analysis and Machine Intelligence, 27(3), Mar. 2005.

[48] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In Proc. of the IEEE, volume 77, pages 257–286, 1989.

[49] E. Mayoraz S. Ben-Yacoub, Y. Abdeljaoued. Fusion of face and speech data for person identity verification. IEEE Trans. on Neural Networks, 10(5):1065–1075, 1999.

[50] I. Verbauwhede S. Yang. Secure fuzzy vault based fingerprint verification system. In Conf. Record of the 38^{th} Asilomar Conf. on Int. Conf. Signals, Systems and Computers, page 577581, 2004.

[51] P. Kenny S. Yin, R. Rose. A joint factor analysis approach to progressive model adaptation in text-independent speaker verification. IEEE Trans. on Speech and Audio Processing, 15(7):1999–2010, Sep. 2007.

[52] J. Odell D. Ollason P. Woodland S. Young, J. Jansen. The HTK Book(for HTK Version 2.1). Cambridge University Press, Cambridge, 1997.

[53] A. Jain U. Uludag, S. Pankanti. Fuzzy vault for fingerprints. In International Conference on Audio- and Video-Based Biometric Person Authentication, pages 310–319, Nov. 1995.

[54] V. P. Kumar V. Ramasubramanian, A. Das. Text-dependent speaker-recognition using one-pass dynamic programming algorithm. In IEEE Int. Conf. on Acoustics, Speech and Signal Processing, volume 1, pages 14–19, May 2006.

[55] S. Renals V. Wan. Speaker verification using sequence discriminant support vector machines. IEEE Trans. on Speech and Audio Processing, 13(2):203–210, Mar. 2005.

[56] T. E. Boult W. J. Scheirer. Cracking fuzzy vaults and biometric encryption. In IEEE Biometrics Research Symposium at the National Biometrics Consortium Conference, Sep. 2007.

[57] T. P. Gleason D. A. Reynolds W. Shen W. M. Campbell, J. P. Campbell. Speaker verification using support vector machines and high-level features. IEEE Trans. on Speech and Audio Processing, 15(7):2085–2094, Sep. 2007.

[58] W. Yau X. Jiang. Fingerprint minutiae matching based on the local and global structures. In Proc. of the 15^{th} Int. Conf. on Pattern Recognition, volume 2, pages 1038–1041, Sep. 2000.

[59] X. Luo T. Zhang Y. He., J. Tian. Image enhancement and minutiae matching in fingerprint verification. Pattern Recognition Letters, 24:1349–1360, Oct. 2003.

[60] M. Carey Y. Liu, M. Russell. The role of dynamic features in text-dependent and -independent speaker verification. In IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pages 669–672, 2006.

[61] R. D. Zilca. Text-independent speaker verification using utterance level scoring and covariance modeling. IEEE Trans. on Speech and Audio Processing, 10(6):363–370, Sep. 2002.