Pontifícia Universidade Católica
do Rio de Janeiro

Abel Sebastián Santamarina Maciá

An Evaluation of Bimodal Recognition Systems Based on Voice and Facial Images

DISSERTAÇÃO DE MESTRADO

Dissertation presented to the Programa de Pós-Graduação em Engenharia Elétrica of the Departamento de Engenharia Elétrica, PUC-Rio as partial fulfillment of the requirements for the degree of Mestre em Engenharia Elétrica.

Advisor: Prof. Raul Queiroz Feitosa

Rio de Janeiro

May 2016

Pontifícia Universidade Católica
do Rio de Janeiro

Abel Sebastián Santamarina Maciá

An Evaluation of Bimodal Recognition Systems Based on Voice and Facial Images

Dissertation presented to the Programa de Pós-Graduação em Engenharia Elétrica of the Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC-Rio, as partial fulfillment of the requeriments for the degree of Mestre.

Prof. Raul Queiroz Feitosa
Advisor
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Alvaro de Lima Veiga Filho
Departamento de Engenharia Elétrica – PUC-Rio

Prof.ª. Karla Tereza Figueiredo Leite
UEZO

Prof. Márcio da Silveira Carvalho
Coordinator of the Centro Técnico
Científico da PUC-Rio

Rio de Janeiro, May 20th, 2016

Abel Sebastián Santamarina Maciá

The author graduated in Telecommunications and Electronics Engineering from Jose Antonio Echeverria Polytechnic Institute in Havana, Cuba, 2011.

Bibliographic Data

Santamarina Maciá, Abel Sebastián

An evalution of bimodal recognition systems based on voice and facial images / Abel Sebastián Santamarina Maciá ; advisor: Raul Queiroz Feitosa. – Rio de Janeiro : PUC-Rio, Departamento de Engenharia Elétrica, 2016.

92 f. : il. color. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2016.

Inclui referencias bibliográficas

1. Engenharia elétrica – Tese. 2. Fusão de escores baseada em densidade. 3. Fusão de escores baseada em transformação. 4. Fusão de escores baseada em classificadores. 5. GMM/UBM. 6. IVector. I. Feitosa, Raul Queiroz. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. III. Título.

CDD: 621.3

Acknowledgments

I deeply thank my advisor, Prof. Raul Queiroz Feitosa, for his support, his advices and talks, his patience and comprehension throughout the development of my dissertation.

I thank PUC-Rio teachers for the formation and knowledge acquired.

I thank CAPES for the financial support.

I thank my parents, Alberto and Lourdes and my brother, Gaston, for their support and unconditional love.

I thank Tamara for her patience and her unconditional support and love.

I thank all my friends at PUC-Rio for sharing their company, friendship and support, and especially my colleagues at LVC for their valuable discussions and advices.

Abstract

Santamarina Maciá, Abel Sebastián; Feitosa, Raul Queiroz (advisor); **An Evaluation of Bimodal Recognition Systems Based on Voice and Facial Images.** Rio de Janeiro, 92p. Master Dissertation - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

The main objective of this dissertation is to compare the most important approaches for score-level fusion of two unimodal systems consisting of facial and independent speaker recognition systems. Two classification methods for each biometric modality were implemented: a GMM/UBM and an I-Vector/GPLDA classifiers for speaker independent recognition and a GMM/UBM and LBP-based classifiers for facial recognition, resulting in four different multimodal combination of fusion explored. The score-level fusion methods investigated are divided in Density-based, Transformation-based and Classifier-based groups and few variants on each group are tested. The fusion methods were tested in verification mode, using two different databases, one virtual database and a bimodal database. The results of each bimodal fusion technique implemented were compared with the unimodal systems, which showed significant recognition performance gains. Density-based techniques of fusion presented the best results among all fusion approaches, at the expense of higher computational complexity due to the density estimation process.

Keywords

Density-based Score fusion; Transformation-based Score fusion; Classifier-based Score fusion; GMM/UBM; I-Vector; LBP.

Resumo

Santamarina Maciá, Abel Sebastián; Feitosa, Raul Queiroz (orientador); **Uma Avaliação de Métodos de Fusão para Sistemas Bimodais de Reconhecimento Baseados em Voz e Imagens Faciais.** Rio de Janeiro, 92p. Dissertação de Mestrado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Esta dissertação tem como objetivo avaliar os métodos de fusão de escores mais importantes na combinação de dois sistemas uni-modais de reconhecimento em voz e imagens faciais. Para cada sistema uni-modal foram implementadas duas técnicas de classificação: o GMM/UBM e o I-Vetor/GPLDA para voz e o GMM/UBM e um classificador baseado em LBP para imagens faciais. Estes sistemas foram combinados entre eles, sendo 4 combinações testadas. Os métodos de fusão de escores escolhidos se dividem em três grupos: Fusão baseada em densidade, fusão baseada em transformação e fusão baseada em classificadores, e foram testadas algumas variantes para cada grupo. Os métodos foram avaliados em modo de verificação, usando duas bases de dados, uma base virtual formada por duas bases uni-modais e outra base bimodal. O resultado de cada técnica bimodal empregada foi comparado com os resultados das técnicas uni-modais, percebendo-se ganhos significativos na acurácia de reconhecimento. As técnicas de fusão baseadas em densidade mostraram os melhores resultados entre todas as outras técnicas, mais apresentaram uma maior complexidade computacional por causa do processo de estimação da densidade.

Palavras Chaves

Fusão de escores baseada em densidade; Fusão de escores baseada em transformação; Fusão de escores baseadas em classificadores; GMM/UBM; I-Vector; LBP.

Content

List of Figures

List of Tables

List of Abbreviations

ASR	Automatic Speaker Recognition
AUC	Area Under ROC Curve
CMC	Cumulative Match Characteristics
CMVN	Cepstral Mean and Variance Normalization
DCT	Discrete Cosine Transform
EER	Equal Error Rate
EM	Expectation-Maximization
FA	Factor Analysis
FAR	False Acceptance Rate
FFT	Fast Fourier Transform
FRR	False Rejection Rate
FRS	Facial Recognition System
GAR	Genuine Acceptance Rate
GMM	Gaussian Mixture Models
GPLDA	Gaussian Probabilistic Linear Discriminant Analysis
ICB	International Conference on Biometrics
IoT	Internet of Things
IR	Infrared
IV	I-Vector
KDE	Kernel Density Estimation
LBP	Local Binary Patterns
LDA	Linear Discriminant Analysis
MAD	Median Absolute Deviation
MAP	Maximum A-posteriori Probability
MFCC	Mel-Frequency Cepstral Coefficients
ML	Maximum Likelihood
MML	Minimum Message Length
MoG	Mixture of Gaussians
NIST	National Institute of Standards and Technology
OA	Overall Accuracy
PCA	Principal Components Analysis
PIN	Personal Identification Number

PLDA	Probabilistic Linear Discriminant Analysis
QLQ	Quadric-Line-Quadric
RBF	Radial Basis Functions
RF	Random Forests
ROC	Receiver Operating Characteristics
SFFS	Sequential Forward Floating Selection
SNR	Signal to Noise Ratio
SVM	Support Vector Machine
TPR	True Positive Rate
UBM	Universal Background Model

1
INTRODUCTION

1.1.
Overview

In recent years, biometric recognition has gained a lot of attention due to the growing necessity for security in our interconnected society. Fingerprint readers, voice recognition and facial identification technologies are now customary in today's personal devices while other techniques like iris scanners, gait tracking and even palm veins scanners are emerging as well in a broad range of applications.

Traditional knowledge-based (passwords, PIN numbers, etc.) or token-based methods (ID-cards, physical keys, etc.) for creating and verifying a person identity, have evidenced to be insufficient for the security standards existing nowadays. Anything "we know" or "we owns" can be lost, manipulated, shared or stolen, presenting a high risk of security breach for the identity. Moreover, these mechanisms cannot provide vital functions like non-repudiation, or multiple instances detection.

With the revolution of personal computing (tablets, smartphones, wearables, Internet of Things (IoT) devices, etc.), it has become increasingly important to develop more reliable identification systems that can provide higher degrees of security and stronger authentication schemes. In this context, biometric recognition has proven to be an excellent solution to the problem of identity determination, since the biometric attributes are inherent to an individual, expressing "who we are", and thus making it very difficult to forget, manipulate or share.

From a general perspective, biometric recognition is the process of identifying a person identity by means of physiological and behavioral characteristics. Examples of those characteristics include face, voice, fingerprint, iris, gait and others, which are referred in the biometric literature as traits, indicators, identifiers or modalities.

Although biometric systems have been successfully implemented in a large amount of applications, with several advantages over the traditional passwords and tokens methods, they also have some limitations, like accuracy, scalability and usability [1]. In this sense, no single biometric is expected to effectively satisfy the requirements of all verification or identification applications.

Most of the current implemented systems are based on one single biometric modality, which are called unimodal biometric systems. Often they have to contend with noisy data (because of the deformable nature of biometric traits, environmental noise, defective sensors, user's accessories occlusions, etc.), non-universality (impossibility to collect meaningful biometric data from a subset of users), inter-user similarity (lack of uniqueness between features of different individuals), intra-user variations (large variations on samples from the same individual making difficult an invariant representation), spoof attacks (deliberate manipulation of one's biometric traits in order to avoid recognition) and unacceptable error rates, factors that make them unsuitable in more constrained scenarios.

Some of these limitations can be overcome by deploying multimodal biometric systems that integrate the evidence presented by multiple sources of information. Using different biometric traits from an individual in a multibiometric system has several advantages. In a first place, there is a clear positive impact on the overall accuracy of the biometric system, as the individuals can be more discriminative in a larger feature space. These systems can also address the problem of non-universality when it is impossible to enroll a person in the system using a specific biometric but can be enrolled using another one. They can also provide certain level of flexibility in specific applications as the users can authenticate using any of their registered biometrics. Other advantage is the reduction of the noise effect in the system, because the lack of quality of one biometric can be compensated with another biometric instance that provides sufficient discriminatory information to make a decision. They are also more robust against spoof attacks because it is more difficult to mimic multiple biometrics simultaneously.

However, multibiometric systems have some drawbacks as well, like a higher computational and storage cost, the additional time required to enroll the users and higher concerns for protecting the biometric models in databases. Therefore it is

important to study these systems in detail before implementing them in real applications.

1.2. Motivation

Research works in the area of classifier combination and specifically in multibiometrics systems are vast [2-13]. Efforts have been oriented in different directions according to the level where the fusion is performed. Among all fusion techniques available, the score-level fusion offers the best trade-off in terms of the information content and the ease in fusion [1]. For these reasons, in this work, it is investigated the score-level fusion in a multimodal biometric system, using voice and face biometrics.

Although bimodal speaker and facial identification systems have been studied in previous works with satisfactory results in the late 1990's and at the early 2000's, in recent days new algorithms for feature extraction and classification of these biometric modalities have emerged, that further improve their accuracy. In this regard, state-of-the-art techniques like *Factor Analysis* (**FA**) and *I-Vector framework* has gained great popularity in the speaker and language recognition community. For face recognition, on the other hand, texture-based methods like *Local Binary Patterns* (**LBP**) has become standardized in automatic face recognition applications when high accuracies are pursued.

Other aspect worth mentioning is the availability of more datasets that contains multibiometric data. In the early studies of multibiometric fusion, there were only few databases containing more than one biometric modality, but nowadays there are available several databases with high amounts of data from diverse biometric traits. Few examples of these datasets are the XM2VTS [14] and BANCA [15] databases, which combine face and speech, BIOMET [16] database that consists of five different modalities (speech, face, hand, fingerprint and signature), BiosecurID [17] that combines eight unimodal biometric traits (speech, iris, face, handwritten signature and handwritten text, fingerprints, hand and keystroking), MOBIO [18] dataset, which is a bi-modal face and speech dataset took from mobile devices, among others.

Finally, techniques and algorithms used for biometric fusion have also developed as well, with a wide list of well-studied and advanced methods of fusion based on density estimation, normalization and classification techniques. These methods will be the subject of study in this dissertation.

1.3. Objectives of the dissertation

The **general objective** of this research is to compare the most important algorithms for score-level fusion in the development of a multimodal biometric system that combines an independent speaker and a facial recognition systems.

Other important **specific objectives** in this work are the following:

- Develop a Text-Independent Automatic Speaker Recognition System (SRS) that provides state-of-the-art performance.
- Compare different speaker and facial recognition algorithms for identification and verification tasks.

One final goal in this dissertation is to create and test a Graphical User Interface (GUI) tool for testing the developed systems and algorithms used in this work.

1.4. Organization of the reminder parts

The following parts of this document are structured as follows:

- Chapter 2 presents an overview of the state of the art in the area of multibiometric fusion, with an emphasis in the score-level fusion in multimodal systems.
- Chapter 3 details the algorithms used in the implementation of the speaker and facial recognition classifiers, as well as for the fusion of biometrics.
- Chapter 4 presents the configuration scheme for both unimodal systems, describing the datasets used, the metrics considered for evaluating the accuracy of independent classifiers and the methodology followed for finding the best unimodal systems configurations. They are presented the

results obtained in the assessment of the individual biometrics classifiers separately.

- Chapter 5 presents the experimental analysis for the biometric combination step and shows the results for each scheme of fusion implemented. The results are discussed in the last part of the chapter.
- Chapter 6 presents the final conclusions of this work, and discusses the future directions that could be taken for further development in this research area

2
RELATED WORKS

This chapter introduces some important concepts in multibiometric systems, such as their main functionalities and the levels and categories of fusion. In addition, examples of the most relevant works of data fusion and combination of classifiers are presented, with emphasis in those related to the score-level fusion.

2.1.
Biometric functionalities

The functionalities or operating scenarios of a biometric system can be classified as verification (authentication) and identification. In verification mode, a user claims for an enrolled identity and the system determines if the claim is true or false, so the query is compared only with the template corresponding to the claimed identity [1]. In practical terms, the biometric classifier provides a match/non-match decision for each verification trial, based on a specified operational threshold. If a claim is accepted it is said to be a "genuine" individual, otherwise it is considered an "impostor".

The identification functionality on the other hand can be subdivided as open-set identification or closed-set identification. In a closed-set identification scheme, for each test sample presented at the input, the system is forced to make a decision in favor to one of the individual identities enrolled in the gallery, based on a similarity or dissimilarity metric. In an open-set identification scheme, an extra option exists, in which the probe sample could not pertain to any of the enrolled identities in the system, and in that case the probe is rejected (i.e., no suitable identification is found on the system) [1].

In Figure 2-1 it is illustrated the structure of a biometric system, consisting of enrollment and authentication stages, for both verification and identification modes.

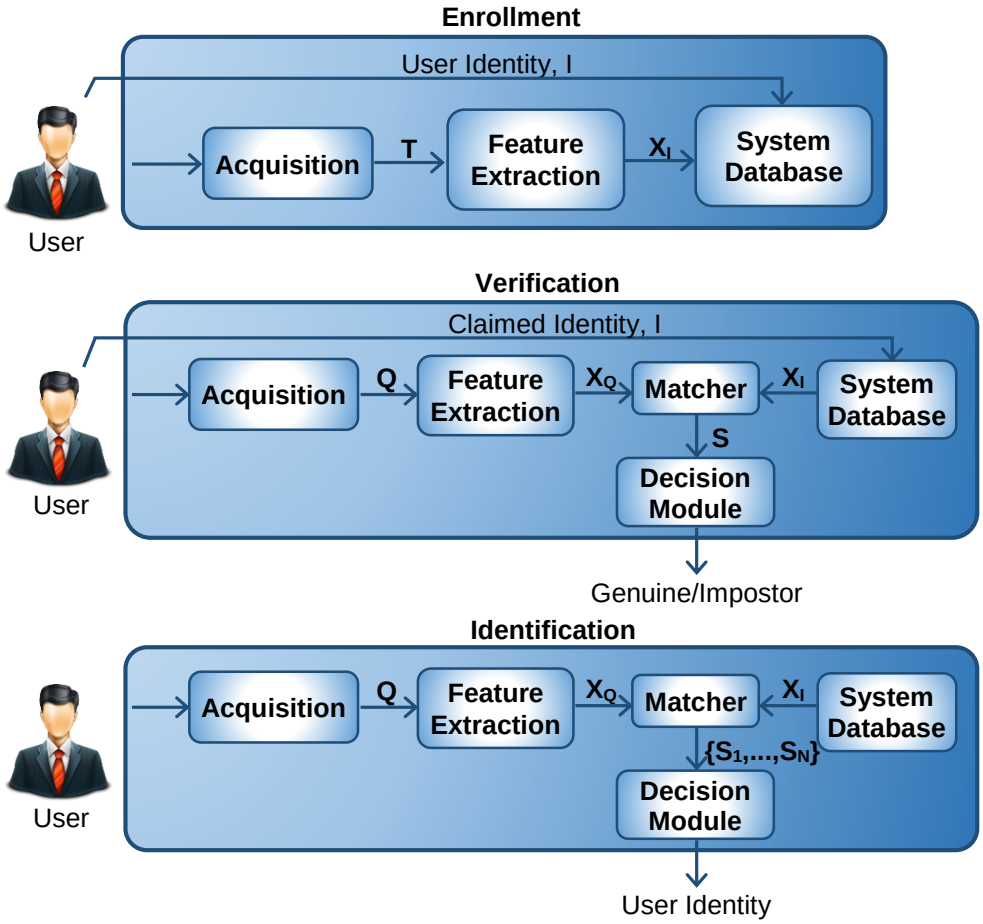

Figure 2-1: Structure of a Biometric System showing the Enrollment and Recognition phases. Here, T represent the sample in the enrollment, Q the query biometric sample during recognition, XI and XQ represent the template and query feature sets respectively, S represents the score and N is the number of users enrolled in the gallery. (Figure modified from [1]).

2.2. Categories and levels of fusion in Multibiometric Systems

According to the literature, multibiometric fusion can be implemented at different levels (levels of fusion) and with different sources of information (categories of fusion) [19]. The categories of fusion define what inputs or processes are being used for fusion, whereas the levels of fusion define how the fusion is performed.

The categories can be divided in six major groups: (1) *Multi-sample*, when the fusion is made among samples from the same source (e.g. face images from a video sequence, different recordings of a speaker, etc.); (2) *Multi-instance*, when multiple instances of the same biometric trait are fused (e.g. fingerprints from

multiple fingers, iris from left and right eyes, etc.); (3) *Multi-modal*, when the fusion is using different biometric modalities (e.g. face, voice, fingerprints, etc.); (4) *Multi-algorithm*, when the results from multiple algorithms that process each individual sample are fused (e.g. Gaussian Mixture Models (GMM), Probabilistic Linear Discriminant Analysis (PLDA), etc.); (5) *Multi-sensor*, when information of the same biometric trait, captured by different sensor types is fused (e.g. optical and capacitive fingerprint sensors); (6) *Metadata*, when external non-biometric information is used to enhance the biometric recognition (e.g. measurement of sample quality, demographic information, etc.)[19].

The levels of fusion are divided in the following groups: (1) *Sensor-level*, the raw information captured by the biometric sensors is fused; requires the biometric sensor to provide compatible inputs (e.g. mean of sequence of images); (2) *Template/feature-level*, multiple features or representations of the biometric data are fused to form a single feature; (3) *Score-level*, multiple samples, instances or modalities are compared and the resulting similarity or distance scores are combined into a single fused score; (4) *Rank-level*, the scores of the system in identification mode are viewed as a ranking of the enrolled identities, ordered in decreasing order of confidence and they are fused using a consensus rank; (5) *Decision-level*, similar to *score-level* but the scores are converted to match/non-match decisions and then fused; typically, in this mode, a consensus strategy is used, like majority voting, logical combination rules, etc. [19].

2.3. Multibiometric fusion

The idea of consolidating biometrical information from multiple sources is not new and it has been extensively studied in the literature [2-13]. Examples of fusion performed at each level and using the categories mentioned before are vast, and it is important to remark that this work does not intend to make an extensive review, but rather cover only the most relevant works related to the objectives of this dissertation.

In the next sections, the two broad classes of fusion, i.e. fusion prior to matching and fusion after matching, are covered. The sensor-level and feature-level

fusion schemes belong to the first class, while score-level, rank-level and decision-level fusion schemes belong to the second.

2.3.1. Fusion prior to matching

Sensor-level Fusion

The first stage in designing a multibiometric system is to determine the information to be combined. In a typical biometric system, the amount of information available to the system gets compressed as one proceeds from the sensor module to the decision module. One initial choice is raw data captured by the sensors, which contains the highest information because no process has been applied to the signal (e.g., the record of an utterance, a face image, a fingerprint, etc.). However, it is probable that the data is contaminated by noise or contains other nuisance effects on it (e.g. reverberation in the speech, non-uniform illumination in facial images, etc.), and therefore in some cases it is necessary to apply filtering or some sort of pre-processing technique to clean the biometric samples before fusion.

One issue with this level of fusion is that, in order to combine the biometric raw data directly from various sensors, they have to be compatible. In other words, this fusion is only applicable if the multiple sources represent samples of the same biometric trait, obtained either using a single sensor or different compatible sensors.

Examples of this approach of fusion can be found in [20] in which multiple 2D face images obtained from different viewpoints were stitched together to form a 3D model of the face. In [21] the authors followed a similar method to perform mosaicking of five views of a face at different angles to create a panoramic face construction in real time.

Another typical application of sensor level fusion is the mosaicking of various partial fingerprints impressions of a person in order to create a better fingerprint image. Examples of this technique have been studied in [22-25] with satisfactory results in terms of performance gains.

Other useful examples of sensor level fusion are discussed in [26-28], where visible and thermal infrared face images at sensor level are fused. By using IR images in conjunction with visible images, illumination challenges in facial

recognition applications can be properly addressed as the IR images are relatively insensitive to illumination changes.

Feature-level Fusion

In this level of fusion there exist two alternatives for combining the features sets extracted from multiple biometric sources, depending on whether they are homogeneous or non-homogeneous. The feature sets are homogeneous when they are obtained by applying the same feature extraction algorithm to multiple samples of the same biometric trait (e.g. minutia sets from two impressions of the same finger). In this case the resulting feature set can be formed as a weighted average of individual feature sets. In contrast, the feature sets are non-homogeneous when they originate from different feature extraction algorithms or from samples of different biometric modalities (e.g. face and hand geometry). In this case the feature sets can be concatenated to form a single feature set, if the feature sets are compatible. Typically, dimensionality reduction schemes based on feature selection or feature extraction mechanisms are applied to obtain a minimal feature set. The key benefit of this is that it enables detection/removal of correlated feature values improving recognition accuracy.

Some works have been published using this fusion scheme. In [29], the authors studied the feature-level fusion applied to three different scenarios: (i) multi-algorithm, combining PCA and LDA coefficients of face, (ii) multi-sensor, where different color channels of the facial image were integrated using LDA features and (iii) multimodal, in which facial and hand features were combined. For each scheme, the feature sets were firstly normalized and then concatenated, forming a high dimensional feature vector. Then the sequential forward floating selection (SFFS) feature selection technique was employed to eliminate redundancy and correlated feature values, thus reducing the dimensionality of the feature vector. The results indicated that feature-level fusion is advantageous in some cases.

In [30, 31], Chibelushi et al. combined voice and outer lip-margin features for person identification, using feature sets concatenation, as well as linear transformation PCA and LDA techniques. In this case, the authors demonstrated that the use of feature level fusion in their system is equivalent to increasing the signal-to-noise ratio (SNR) of the audio signal, under adverse acoustic conditions.

A similar idea was used in [32], in which the performance of a system using the correlation between audio-visual features during speech was evaluated. On this work a concatenation of voice and facial features was employed, and the authors reported lower error rates and higher robustness against replay attacks when compared to audio-only, video-only and audio-visual systems which assume audio and visual data to be independent.

Other works includes fusion of face and iris features [33-36], hand geometry and palmprint [37], face and palmprint [38], face and gait to recognize individuals at distance in video [39], among others.

Integration at the feature level has proved in the literature to be challenging for some reasons, specifically: (1) incompatibility in the feature vectors extracted from different modalities, (2) high dimensionality when the concatenation of multiple vectors is carried out, causing the *curse-of-dimensionality* problem [40], where the classification accuracy actually degrades with the addition of new features due to the limited number of training samples, (3) the relationship between the feature spaces of different biometric systems may be unknown and sometimes the features sets can be highly correlated, (4) most commercial biometric system vendors do not provide access to the feature sets. These constraints have led this fusion scheme to limited success and relatively less attractiveness versus other fusion schemes, like score-level fusion.

2.3.2. Fusion after matching

Score-level Fusion

This is the most commonly used method for integrating information in multibiometric systems. The fusion at this level is also known as measurement level or confidence level, and it combines the match scores output by the classifiers in order to make a recognition decision. The match scores can represent a similarity or dissimilarity (distance) metric between the input and the template biometric feature vectors, and they can have different ranges and different probability distributions. These and other difficulties have motivated the research on this fusion scheme.

The related works in this level of fusion can be broadly classified into three groups: transformation-based score fusion, density-based score fusion and classifier-based score fusion, although there are other variations, including the combination of these groups. The study of these schemes of fusion are the center of this work, and they will be explained in detail in subsequent chapters.

The first works using score level fusion were proposed at mid ‘90s, and they focused mainly in the fusion of voice and facial biometric traits. In [6], Brunelli and Falavigna proposed two fusion approaches for combining scores from acoustic and visual features from a non-public database. In the first approach, they fused the outputs of two speech classifiers and three face classifiers, using a Tanh normalization, which relies on a robust estimation of location and scale parameters of score distributions, in combination with a geometric average. For the second approach, they proposed a hybrid rank/score fusion using a HyperBF network in a classification-based scheme.

In [41], Duc et al. proposed a bimodal fusion of face and speech experts from M2VTS dataset [42], for person authentication using simple averaging of scores and a more sophisticated Bayesian integration scheme presented by Bigün et al. in [2].

In [4], Kittler et al. developed a theoretical framework considering the biometric multimodality as a classifier combination problem. In this work, the authors compared fusion schemes like sum rule, product rule, max rule, min rule, median rule and majority voting under a probability Bayesian perspective. The scores of frontal face, face profile, and voice experts from M2VTS dataset, were converted into posterior probabilities and then fused using the aforementioned fusion schemes. They concluded that the best combination method was the simple sum rule of posterior probabilities after testing the sensitivity of all fusion methods to estimation errors. In [43], Verlinde et al. also used the Bayesian approach with the same biometric modalities and the same database. The scores in this case were converted into posterior probabilities assuming they follow a Gaussian distribution and the mean and variance were estimated from the training data. They also considered a Logistic Regression technique for inferring the posterior probabilities. Finally the scores were fused using the product, assuming they were independent from one another.

Following the Bayesian approach presented in [4] and [43], the National Institute of Standards and Technology (NIST) conducted an study [44] where eight biometric fusion techniques were compared. They concluded that the Product of Likelihood Ratios gives consistently the most accurate results for score-level fusion, but it is the most complex method to implement as it requires the explicit estimation of the match score densities; moreover, the fusion heavily depended on the reliability of the density estimation process, in this case the kernel density estimator (KDE) used. Nandakumar et al. proposed in [45] the use of Gaussian Mixture Models (GMM) for estimating the probability density functions, which is easier to implement than KDE and models quite effectively the scores densities. They also showed that quality measurements of the biometric samples improved the accuracy and this inclusion should be evaluated in the fusion process whenever this information is available.

In relation to classifier-based approaches, Verlinde and Chollet in [11, 46], considered the multimodal fusion as a pattern classification problem. Under this point of view, the scores given by individual expert modalities are considered as input patterns to a classifier. They compared the following pattern classification techniques with the results from [43]: Maximum a Posteriori Probability (MAP), k-Nearest Neighbors classifiers, Multilayer Perceptrons, Binary Decision Trees, Maximum Likelihood (ML), Quadratic classifiers and Linear classifiers. They concluded that every fusion method improved the performance over the best single expert and the Logistic Regression offered the best results.

Another classification-based approach was studied in [47], involving Support Vector Machines (SVM) to combine face, fingerprint and on-line signature biometric modalities from MCYT [48] and XM2VTS [14] datasets. A similar SVM classification scheme was used in [49] for fusing iris and face biometric traits from ORL [50] face image and UBIRIS [51] iris databases respectively, using a double sigmoid function to normalize the scores after the matcher's outputs. In [52], an ensemble of classification trees called Random Forests [53] was used for combining fingerprint, face and hand geometry biometrics from a non-public database. In [54], three methods of fusion were employed: (1) a weighted sum of scores, (2) a Fisher's discriminant analysis and (3) a neural network with radial basis function (RBFNN), in a multimodal system consisting of face and iris modalities from a composite of databases.

Other works addresses the transformation-based approach, such as [55], in which several normalization and fusion techniques were tested to fuse the scores from face, fingerprint and hand geometric matchers. Specifically the Min-Max, Decimal Scaling, Z-Score, Median and Median Absolute Deviation (MAD), Double Sigmoid, Tanh estimators and Parzen normalization techniques were used, along with the sum, min and max of normalized scores for fusion. The experiments showed that the Min-Max, Z-Score and Tanh normalization schemes followed by a simple sum of scores results in better recognition performance compared to other methods. They also compared user-specific weights versus common weights to multiple biometric traits of all users, revealing that the former approach is advantageous over the latter.

Another extensive work comparing several normalization and fusion schemes, was proposed in [56], using fingerprint and face biometrics. The authors also concluded that Min-Max normalization with a simple sum offered the best results when dealing with open populations (e.g. airports), whereas for closed populations (e.g. office environments) an adaptive normalization scheme proposed called Quadric-Line-Quadric (QLQ) with user-specific weighted sum provides the best accuracy.

Rank-level Fusion

This level of fusion is used in identification systems, where the scores are sorted in a ranking form and then fused using consensus strategies.

Ho et al. in [9] describe three methods to combine the ranks assigned by different matchers, specifically, the highest rank method, the Borda count method and the logistic regression method.

Decision-level Fusion

The Decision-level fusion is also known as abstract fusion because it is used when there is access only to the decisions taken by the individual classifiers. It is the easiest fusion level among the others and is the less studied in literature, as it is often considered inferior to matching score-level fusion, on the basis that decisions are too ‘‘hard’’ and have less information content compared to ‘‘soft’’ matching scores [57].

The methods proposed in the literature for this level of fusion include logical combination rules, like "AND" and "OR" rules [58], majority voting [59], weighted majority voting [60], Bayesian decision fusion [61], the Dempster-Shafer theory of evidence [61] and behavioral knowledge space [62].

3
THEORETICAL FUNDAMENTALS

In this chapter, it is presented the general structure of the multimodal system build in this study, describing the algorithms used in the unimodal systems as well as the techniques used in the biometric fusion.

3.1.
General Dataflow Scheme

In the figure below (Figure 3-1), it is shown the proposed bimodal fusion scheme for combining the unimodal biometric systems.

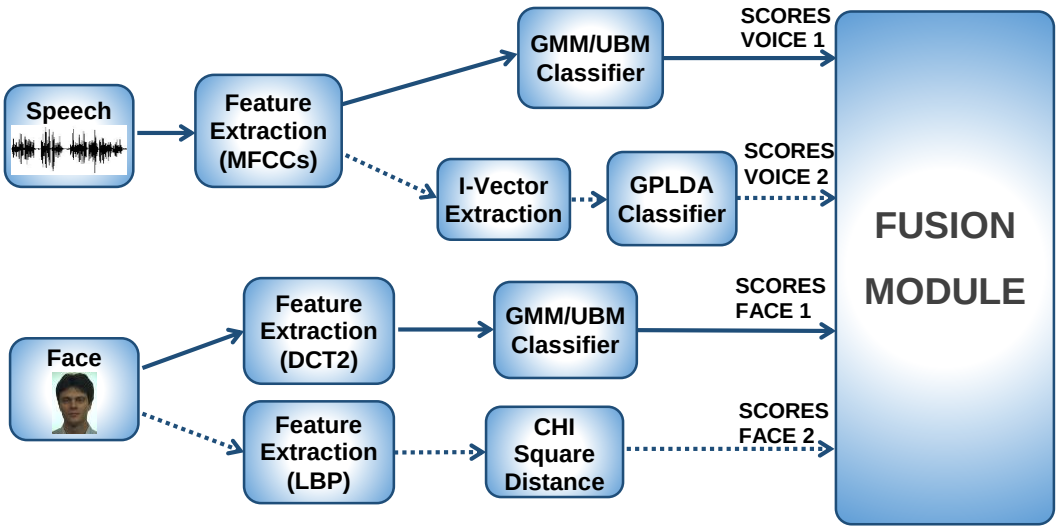

Figure 3-1: Unimodal systems schemes and algorithms.

As it can be observed, each biometric system has in common a feature extraction step, where salient characteristics of voice and face biometrics are summarized by compact representations. This step is followed by a classification stage, where two different methods for each biometric modality are tested. At the end, a vector of scores is generated at each matcher output for the defined test individuals in the dataset.

In the speaker recognition system, we used the Mel Frequency Cepstral Coefficient (MFCC) features, widely adopted in speaker recognition tasks [63, 64]. Next, for the back-end, two different approaches were implemented: i) the well-known GMM/UBM classifier [65] and ii) the I-Vector/GPLDA classifier framework.

For the face recognition system, two different approaches were tested as well. Initially, a similar structure to the first speaker system was used, using as the front-end the 2D Discrete Cosine Transformation (DCT) and the GMM/UBM matcher for the back-end. The other method of interest, as shown in Figure 3-1 was the Local Binary Patterns (LBP), which is a well-known texture-based technique that has become increasingly popular for face recognition. This "package" is completed with the Chi Square Distance Classifier.

In the following sections each of these algorithms and approaches will be covered in detail.

3.2. Closed-Set Text-Independent Speaker Recognition System

In this work, a closed-set text-independent speaker recognition system was implemented. A closed-set system assumes that every test user claims for an identity that is already in the set of enrolled speakers. Text-independent recognition means that the users are not compelled to speak any particular text or speech, so the system does not pose any constraint to the linguistic content of the speeches. In addition, two modes of operation were tested, this is, identification and verification.

3.2.1. Data Acquisition and Preprocessing

This subsection describes superficially the procedures adopted in this work for data acquisition and preprocessing of speech in the process of feature extraction. A more detailed description of the algorithms described hereafter can be found in [63, 64].

In Figure 3-2 it is shown the processing chain to transform the speech utterances into MFCC feature vectors. The first step involves a high-pass pre-

emphasis filter to emphasize the high frequencies and compensate the human speech production process which tends to attenuate the high frequencies. Following, the temporal signal is divided in "chunks" (called frames) of 25ms of duration with 10ms of increment. Each frame is multiplied by a windowing Hamming function that smooths the borders. A Fast Fourier Transform operation is then applied to each frame, yielding complex spectral values. The phase of the FFT is discarded and only the magnitude is considered. Later, a filterbank of 40 triangular filters (K=40) is constructed in the Mel-Scale, which is a logarithmic scale in frequency domain that is perceptually more meaningful for humans [63, 64], and then these filters are multiplied by the FFT coefficients, reducing the total FFT magnitude coefficients to a more compact representation. The output of the Mel-filters is transformed into the logarithmic domain, and then projected into an orthogonal Discrete Cosine Transformation (DCT) basis. In practice, the first 13 DCT coefficients are preserved.

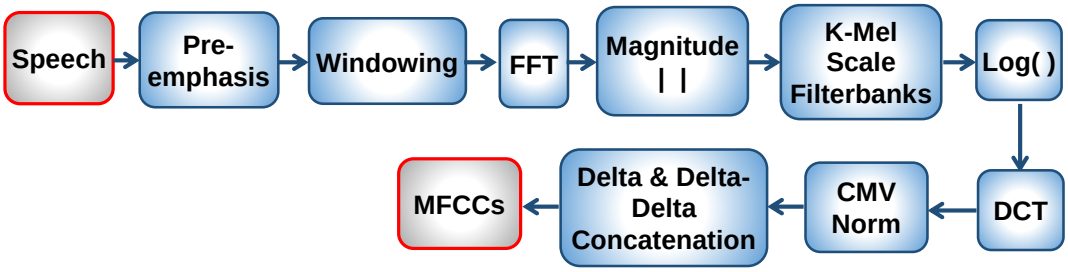

Figure 3-2: Speech signal processing chain for MFCCs computation.

Next, the coefficients are mean and variance normalized (Cepstral Mean and Variance Normalization – CMV Norm), and then they are concatenated with the delta and delta-delta temporal derivatives computed over adjacent frames (normally a span of 2 frames from the left and right). Thus, it ends up with a vector of dimension 39 (13 Cepstral coeff. + 13 Delta coeff. + 13 Delta-Delta coeff.).

3.2.2. Classification

Gaussian Mixture Models (GMM) and GMM/UBM.

Gaussian mixture models (GMMs) were first introduced as a method for speaker recognition in the early 1990's and have since then become the de-facto reference method in speaker recognition [65, 66]. A Gaussian Mixture Model

(GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system. Its parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm [67] or Maximum A-Posteriori Probability estimation (MAP) from a well-trained prior model.

The weighted sum of M-component Gaussian densities is given by the equation:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i \, g(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \qquad \textbf{(3-1)}$$

where $\mathbf{x}$ is a *D*-dimensional continuous-valued data vector (i.e. measurement or features), $w_i, i = 1,2, \dots, M$, are the mixture weights, and $g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1,2, \dots, M$, are the component Gaussian densities. Each component density is a *D*-variate Gaussian function of the form:

$$g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)' \, \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\right\} \qquad \textbf{(3-2)}$$

with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. The mixture weights satisfy the constraint $\sum_{i=1}^{M} w_i = 1$, being all non-negative.

The mean vectors, covariance matrices and mixture weights from all component densities parameterize the complete Gaussian mixture model, represented by the notation:

$$\lambda = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}, \qquad i = 1, \dots, M \qquad \textbf{(3-3)}$$

The classification using this method can be seen as a likelihood ratio (LR) detector [65] between a GMM model of a given speaker and a model that describes the entire space of possible alternatives to that speaker. Formally,

$$\Lambda(\mathbf{x}) = \log p\left(\mathbf{x}\middle|\lambda_{hyp}\right) - \log p\left(\mathbf{x}\middle|\lambda_{\overline{hyp}}\right) \qquad \textbf{(3-4)}$$

where the likelihood ratio is converted into a difference of log-likelihoods between the models.

In practice, this alternative model $p\left(\mathbf{x}\middle|\lambda_{\overline{hyp}}\right)$, common to each hypothesized speaker, is constructed by pooling data from several speakers and then training a single model, called Universal Background Model (UBM), using the Expectation-Maximization algorithm [67]. Once trained the UBM, the GMM speaker models are constructed with a form of Bayesian adaptation, called Maximum A-Posteriori

Probability (MAP) estimation. Although it is possible to adapt all the parameters of the UBM, it is customary only to adapt the means, as it has been empirically proven to offer the best results [65]. Figure 3-3 shows this idea in a synthetic 2-dimensional space.

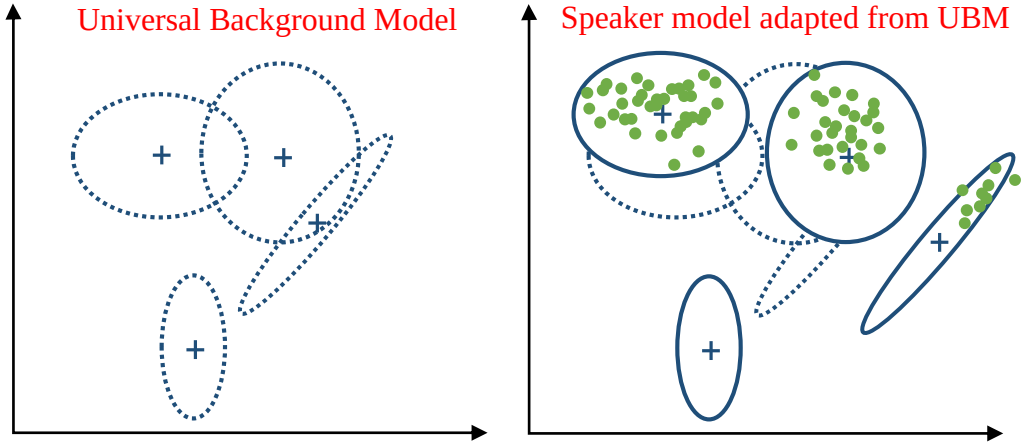

Figure 3-3: MAP algorithm used to adapt the means of the UBM based on the observed data from speaker.

I-Vector framework

In order to create a better model for the speakers, that takes into account the session variabilities and other nuisances existing in the speaker phones, a new framework was developed by Kenny et al. in [68], with the use of Factor Analysis techniques. This new framework is based on the concept of "supervector" and provides a new fixed-length representation of the variable-length utterances from the users. Formally speaking, given a sequence of N MFCC frames, $O = \{o_t\}_{t=1}^{N}$ with $o_t \in \mathbb{R}^D$ and a UBM model $\lambda_{UBM} = (\{w_k\}, \{\boldsymbol{\mu}_k\}, \{\boldsymbol{\Sigma}_k\})$ with K components, the zero and first-order Baum-Welch statistics are extracted, and the supervector $\theta = \{\theta_1^T, \dots, \theta_K^T\}^T$ is constructed by appending the zero and centered first-order statistics for each component mixture to form a high dimensional vector of dimension KD x 1, as described in [69].

This supervector, which can be perceived as the container for the main differences between users, is assumed to obey an affine linear model of the form:

$$\theta = m + \mathrm{T}x \qquad (3\text{-}5)$$

where $m \in \mathbb{R}^{KD}$ is the mean supervector coming from the UBM, $\mathrm{T} \in \mathbb{R}^{KD \times F}$ is a rectangular matrix of low rank called Total Variability Space, representing the speaker-specific information along with the session variabilities, and $x \in \mathbb{R}^F$ is a

standard normally distributed latent variable called total factors, used to compute the i-vectors. Their dimension can be chosen by the user, being F = 400 a typical value which has empirically offered good results for the speaker recognition problem.

The T matrix is obtained from a set of training feature vectors as described in [70], using the UBM model to compute the required statistics over the training vectors. Once calculated the matrix T, they are computed the total factors for every utterance, and the i-vectors are obtained, following a procedure detailed in [69].

After obtaining the i-vectors for every speaker utterance, a Linear Discriminant Analysis (LDA) algorithm is used to annihilate undesired variabilities and to increase the discrimination between speaker subspaces[69]. This new algorithm further reduces the dimensionality of the data, up to 200. Finally, and before the classification using the *GPLDA* probabilistic generative model, the data is centered (mean normalized), and length normalized.

Gaussian Probabilistic Linear Discriminant Analysis (GPLDA) classifier

Probabilistic Linear Discriminant Analysis (PLDA) was firstly introduced in face recognition in [71] as a technique for separating the within-individual and between individual variations, and then it was successfully adapted to speaker recognition tasks, representing nowadays the state-of-the-art in the field. Two basic assumptions are taken into consideration in this version of the model: i) the speaker and channel components are statistically independent and ii) they are Gaussian distributed. By making these assumptions, the likelihood ratios can be obtained in a closed-form.

The GPLDA model used in the speaker recognition context assumes that the descriptors, in our case the i-vectors, are made of speaker-specific components and undesired variability components (session variabilities or channel components). Formally, given R utterances from speaker i, and denoting the collection of i-vectors as $\{\eta_{i,r}\}$, with $r = 1, \dots, R$, the observed i-vector $\eta_{i,r}$ can be decomposed as:

$$\eta_{i,r} = m + \phi\beta_i + \Gamma\alpha_{i,r} + \epsilon_{i,r} \quad \textbf{(3-6)}$$

where, $m + \phi\beta_i$ describes the between-speaker variability, and $\Gamma\alpha_{i,r} + \epsilon_{i,r}$ describe the channel components. In particular, the columns of ϕ and Γ provide the basis for the speaker-specific subspace (eigenvoice) and the channel dependent subspace (eigenchannel) respectively; the vectors β and α are the latent vectors of these subspaces respectively; ϵ represents the residual components not described by the previous terms, assumed to follow a Gaussian distribution with zero-mean and diagonal covariance $\mathbf{\Sigma}$.

A further simplification is commonly applied, proposed in [72, 73], in which the eigenchannels are discarded and a full covariance matrix is considered on the residual term. In this way, the modified GPLDA is simplified as follows:

$$\eta_r = m + \phi\beta + \epsilon_r \quad \textbf{(3-7)}$$

The model parameters $\{m, \phi, \mathbf{\Sigma}\}$ are obtained from a large collection of development data using the EM algorithm, as described in [71].

For classification using the PLDA mechanism, two i-vectors η_1 and η_2 are presented to the system and the objective is to determine whether both i-vectors share the same latent variable β (hypotheses $\mathcal{H}_s$) or they were generated using different latent variables β_1 and β_2 (hypotheses $\mathcal{H}_d$). The score is computed as the log-likelihood ratio of posterior probabilities between the hypotheses, as follows:

$$score = \log \frac{p(\eta_1, \eta_2 | \mathcal{H}_s)}{p(\eta_1 | \mathcal{H}_d) p(\eta_2 | \mathcal{H}_d)} \quad \textbf{(3-8)}$$

This score is computed between each test-gallery pair of users defined in the testing process.

The general workflows for the two classification methods presented so far are depicted in Figure 3-4 and Figure 3-5. Common to both models is the initial step of training the UBM, in which a training feature set is used, called development in the figure. For the GMM approach, once the UBM is trained, a new set of features called enrollment set is used to create the gallery models (Adapted speakers), using MAP adaptation.

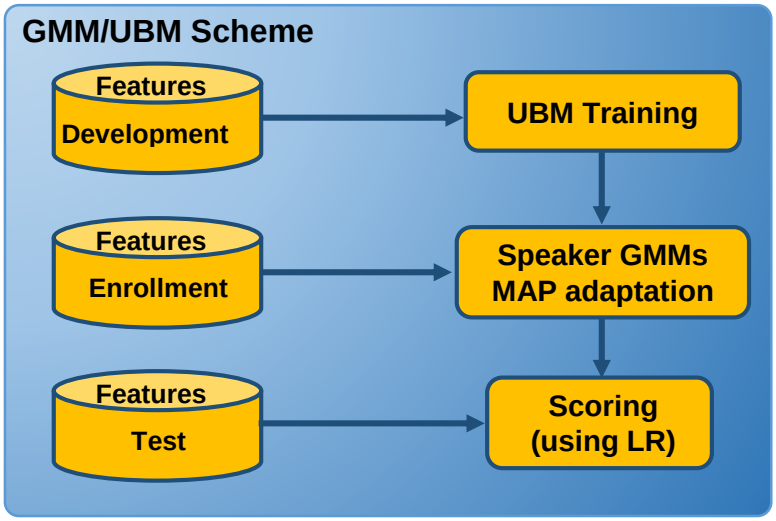

Figure 3-4: Signal flow in the GMM/UBM classification approach.

Finally, every test sample in the test set is scored with the likelihood ratio detector with all gallery models in identification mode and with the claimed identity model in verification mode.

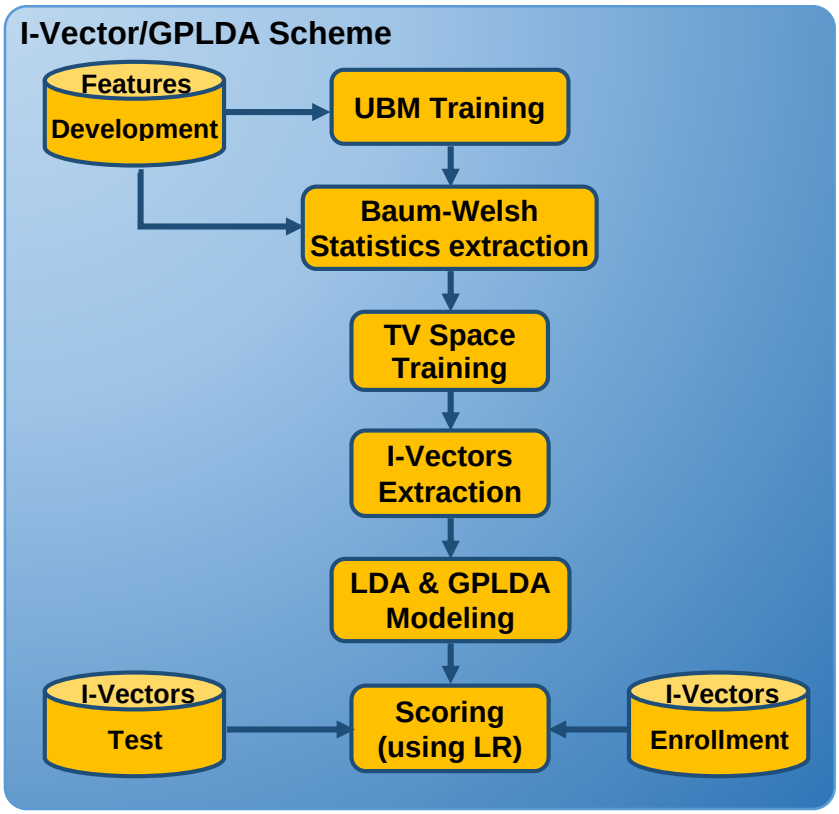

Figure 3-5: Signal flow in the I-Vector/GPLDA classification approach.

For I-Vector/GPLDA scheme, once the UBM is created, the Baum-Welsh statistics are extracted from the training set, and they are used to train the Total Variability Space. Once this space is trained, the I-Vectors are extracted and the Gaussian PLDA Model is created from the training set. The hyperparameters of the

GPLDA model, as well as the variability matrix for the LDA are learned from these training I-Vectors.

Finally, every new user that is to be enrolled to the system has its utterance converted to an I-Vector and stored. Every new test sample repeats the same enrollment process, but it is not saved, just scored with each existing model.

3.3. Facial Recognition System

This section describes the front-end and back-end algorithms used in this work for facial recognition.

3.3.1. Data Acquisition and Preprocessing

For classifying faces the first step is to normalize the images in the database, in order to extract the main characteristics in form of feature vectors. The normalization is composed of two step: a geometric normalization and a photometric normalization. In the present dissertation, the geometric normalization frame the faces to fixed sized boxes of 80x64 pixels, with the eyes coordinates in fixed locations. Then, the photometric normalization compensates for the illumination, contrast, gamma and other effects. This procedure is illustrated in Figure 3-6.

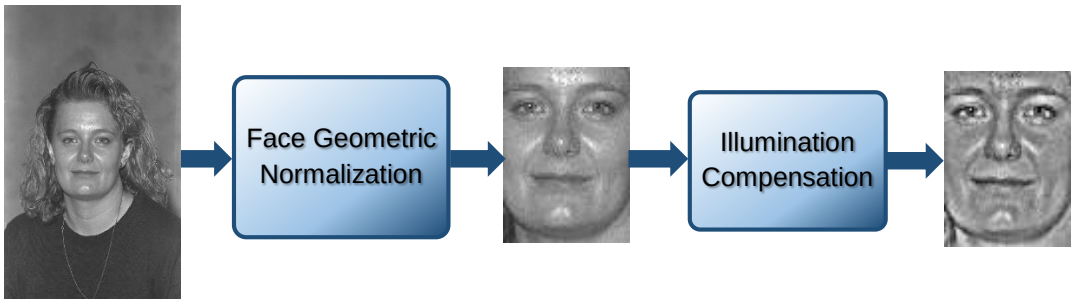

Figure 3-6: Face Geometric and Photometric Normalization.

Two different feature extraction techniques were used, once the images were normalized. In first place, for the GMM/UBM classifier, the normalized images were divided in patches of 16x16 pixels with 50% of overlap between consecutive patches. Then a Discrete Cosine Transformation (DCT-mod2 presented in [74]) was applied to each patch, selecting the first 64 components corresponding to the lowest

frequencies. Finally, the DCT components extracted from each patch are stacked together forming the feature vector of the images. The other approach followed is the Local Binary Patterns (LBP) proposed in [75], and it will be explained in detail below.

Local Binary Patterns

The LBP operator is one of the most used texture-based descriptors in facial recognition applications involving frontal images with minor variations of facial expressions[76]. It is computationally efficient and highly discriminative.

The descriptor is formed by thresholding each pixel intensity of an image at location $\mathbf{x} = \{x, y\}$ with its neighbors at a distance R, as indicated in Figure 3-7. If the sampling point doesn't fall in the pixel's center, a bilinear interpolation is used. The LBP code is formed from the concatenation of the *m* 0's and 1's in an arbitrary but fixed order.

Once obtained the labeled LBP image, it is divided in equally-sized non-overlapping blocks or regions, and a histogram of 59 bins is computed over each block. Each histogram is weighted according to the region it belongs to, so relative importance can be given to some regions in the face over others. The resulting histograms are concatenated in a feature vector that represents the image.

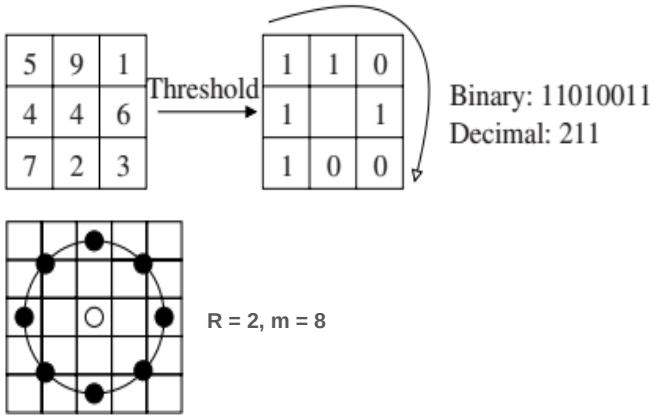

Figure 3-7: The basic LBP operator (taken from [77]).

3.3.2. Classification

The GMM/UBM approach followed in the Facial Recognition System was the same used for the Speaker Recognition counterpart (refer to section 3.2.2).

The matching procedure followed in the LBP approach is based on a dissimilarity measure between weighted histograms from two given images, using a Chi-Square distance function, ruled by the following equation:

$$\chi^2_\omega(\mathbf{S}, \mathbf{M}) = \sum_{i,j} \omega_j \frac{(\mathbf{S}_{i,j} - \mathbf{M}_{i,j})^2}{\mathbf{S}_{i,j} + \mathbf{M}_{i,j}} \quad \textbf{(3-9)}$$

where ω_j is the weight associated with region j; $\mathbf{S}_{i,j}$ and $\mathbf{M}_{i,j}$ represent the feature vectors extracted from a pair of test and gallery images respectively.

3.4. Multibiometric Fusion

In the following sections we describe the algorithms applied in this study for biometric fusion. In the following text we use the word “matcher” to denote a unimodal biometric scheme comprising a feature extraction and a classification model.

3.4.1. Score Fusion Techniques

The Score Fusion methods integrate the information coming from the matcher’s outputs in form of scores. The available techniques are commonly divided in three categories: 1) *Transformation-Based Score Fusion*, 2) *Density-Based Score Fusion* and *Classifier-Based Score Fusion*. In the following sections they will be explained in detail.

It is worth pointing out that a gallery-aggregated score convention will be used. It states as follows: given a database, let G be any enrolled user in the gallery and Q be an identity in the probe set, a score is said to be a genuine score if $Q = G$, otherwise it is an impostor score. In other words, any comparison between an individual sample in the test set and a model of himself in the gallery is labeled as a genuine score and otherwise it is labeled as impostor.

3.4.1.1. Transformation-Based Score Fusion

This method of fusion is based on the combination of scores by simple arithmetic operations like the average sum, product, min and max of the scores.

Fusion

Let's denote s_j^i the i^{th} score output by the j^{th} matcher ($i = 1, \dots, N, j = 1, \dots, R$, N is the number of users enrolled in the gallery at each matcher and R is the number of matchers) and let f^i be the fused score, which can be calculated in a number of different ways, for instance:

- The average sum of scores:

$$f^i = \frac{1}{R}\sum_{j=1}^{R} s_j^i \quad \textbf{(3-10)}$$

- The product of scores:

$$f^i = \prod_{j=1}^{R} s_j^i \quad \textbf{(3-11)}$$

- The minimum of scores:

$$f^i = \min(s_1^i, s_2^i, \dots, s_R^i) \quad \textbf{(3-12)}$$

- The maximum of scores:

$$f^i = \max(s_1^i, s_2^i, \dots, s_R^i) \quad \textbf{(3-13)}$$

The direct fusion of scores is not suitable when the scores are not compatible, i.e. the scores can be either measures of similarity or distance, or have different scales. In this case, a normalization step is needed to transform the scores into a common domain before the fusion. The normalization schemes studied in this work are the Min-Max, the Z-Score and the Tanh normalization.

Normalization

Let's denote ns_j^i the normalized score, μ_j and σ_j the arithmetic mean and standard deviation of the training scores respectively:

- The *Min-Max* normalization is defined as:

$$ns_j^i = \frac{s_j^i - \min_{i=1}^{N} s_j^i}{\max_{i=1}^{N} s_j^i - \min_{i=1}^{N} s_j^i} \quad \textbf{(3-14)}$$

– The *Z-Score* normalization is defined as:

$$ns_j^i = \frac{s_j^i - \mu_j}{\sigma_j} \quad \textbf{(3-15)}$$

– The *Tanh* normalization is defined as:

$$ns_j^i = \frac{1}{2}\{\tanh\left(0.01\frac{s_j^i - \mu_j}{\sigma_j}\right) + 1\} \quad \textbf{(3-16)}$$

The Min-Max and Tanh normalization methods transform the scores into a common range $\{0,1\}$. Using these methods, the scores from the distance classifiers can be converted into similarity scores by subtracting the normalized score from 1. Moreover, the Z-Score method centers the distribution of the normalized scores and equals the variance to 1. These methods have in common that they require a training set to compute their parameters.

3.4.1.2. Density-Based Score Level Fusion

This is the most principled approach for biometric fusion [1] and it is based on estimating the probability density functions of the scores of each class, as explained in the course of this section.

According to the Bayesian decision theory, given an input pattern X composed by the feature vectors derived from R biometric modalities ($X = (x_1, \dots, x_R)$), and $\{\omega_1, \dots, \omega_M\}$, being the possible classes for classification, the input pattern should be assigned to the class ω_r that maximizes the posterior probability, i.e.,

Assign $X \rightarrow \omega_r$ if

$$P(\omega_r|x_1, \dots, x_R) \geq P(\omega_k|x_1, \dots, x_R) \quad \textbf{(3-17)}$$

This rule is called the minimum error-rate decision rule and it assigns no cost to a correct decision and a unit cost to a misclassification error, or in other words, all the errors are equally costly. In real applications, different costs are assigned to the errors. Specifically, if $\eta = \lambda_1/\lambda_2$ is the ratio between the costs associated with false acceptance (λ_1) and false rejection (λ_2) errors respectively, the Bayesian decision rule takes the form:

Assign $X \rightarrow \omega_r$ if

$$\frac{P(\omega_r|x_1, \dots, x_R)}{P(\omega_k|x_1, \dots, x_R)} \geq \eta \quad \textbf{(3-18)}$$

The output of the biometric matchers are scores and not joint posterior probabilities, so in order to apply the Bayesian approach in equation (3-18), it is necessary to transform the independent scores at the matcher's outputs into joint posterior probabilities, or using Bayes Rule, into a joint likelihood ratio, as it will be seen subsequently.

In [43], Verlinde et al. proposed that the match score s_k of the k^{th} matcher is related to its marginal posterior probability by the equation:

$$s_k = f\{P(\omega_k|x_j)\} + \beta(x_j) \quad \textbf{(3-19)}$$

where $f\{\}$ is a monotonic function and $\beta(x_j)$ is the estimation error, which depends on the feature vectors.

In this formula, assuming β is zero, it is reasonable to approximate $P(\omega_k|x_j)$ by $P(\omega_k|s_j)$, which is the posterior probability of class ω_k given the score s_j. Using the Bayes Rule, and assuming that all classes are equally probable, the posterior probability ratio between classes ω_i and ω_k can be stated as follows (also known as density function ratio or likelihood ratio):

$$\frac{P(\omega_i|s_j)}{P(\omega_k|s_j)} = \frac{p(s_j|\omega_i)}{p(s_j|\omega_k)} \quad \textbf{(3-20)}$$

This formula can be extended to joint densities of R matchers, following the same considerations as with previous equation, which yields the Neyman-Pearson theorem, described in [78]:

Assign $X \rightarrow \omega_i$ if

$$\frac{P(\omega_i|s_{1,} \dots, s_R)}{P(\omega_k|s_{1,} \dots, s_R)} = \frac{p(s_{1,} \dots, s_R|\omega_i)}{p(s_{1,} \dots, s_R|\omega_k)} \geq \eta \quad \textbf{(3-21)}$$

where η is the ratio of error costs described in equation (3-18).

Finally, as it can be seen in the equation (3-21), the idea is to estimate the joint density functions of all matchers for each class and then apply the Neyman-Pearson rule. This problem can be addressed in two ways:

1) Estimate the marginal densities functions $p(s_j|\omega_i)$ of the scores of each individual matcher $s_j, j = 1 \dots, R$ for every class $\omega_i, i = 1 \dots, M$, and then use the methods proposed in [4] by Kittler et al. to combine the marginals. These methods

consist on the sum, product, min, max and median of the density functions based on the assumption of statistical independence between them (i.e. the biometric matchers). In this work, only the product and sum of marginal densities were evaluated.

2) Directly estimate the joint density functions of the scores of all matchers that pertain to a given class ω_k (i.e. $p(s_{1,} \dots, s_R|\omega_k)$).

In the case of verification, only the classes *genuine* and *impostor* exist. Therefore, the equation (3-21) can be simplified as follows:

Assign $X \rightarrow genuines$ if

$$\frac{p(s_{1,} \dots, s_R|genuines)}{p(s_{1,} \dots, s_R|impostors)} \geq \eta \qquad \textbf{(3-22)}$$

Density estimation can be also divided in two categories: parametrical or non-parametrical. In the first case, the scores are assumed to follow a known density function (e.g. Gaussian) and the parameters are estimated from the training scores. For the second case, no prior assumption is made about the form of the density function, and the estimation is data-driven. In this work, two non-parametric methods were used. For the approach in **1)**, the Kernel Density Estimation algorithm was employed, whereas for the approach in **2)**, the Mixture of Gaussians algorithm was used.

Kernel Density Estimation

The KDE algorithm is a well-known non-parametric technique for estimating the probability density function followed by a set of data. It is ruled by the following equation (also known as Parzen window estimator):

$$\hat{f}(s) = \frac{1}{hN}\sum_{i=1}^{N} K(\frac{s - s_i}{h}) \qquad \textbf{(3-23)}$$

where K is the kernel function, which satisfies $\int_{-\infty}^{\infty} K(x)dx = 1$; s_i is the observed score vector, $i = 1{:}\,N$ (N being the vector size); h is the bandwidth of the kernel. It is common to select a kernel symmetrical about zero, and for this work, it was selected a Gaussian kernel. The other critical parameter is the bandwidth, which was chosen empirically according to the expression:

$$h = \frac{\sigma}{log(N)} \qquad \textbf{(3-24)}$$

where σ is the standard deviation of the training match score vector for a given class.

GMM-based Density Estimation

The GMM algorithm presented in 3.2.2 can be used as well for estimating the multivariate conditional density function of scores of R matchers. For genuine and impostor classes in verification mode, the estimates of the density functions $f_{gen}(\boldsymbol{s})$ and $f_{imp}(\boldsymbol{s})$ are obtained as a mixture of Gaussians as follows:

$$\hat{f}_{gen}(\boldsymbol{s}) = \sum_{i=1}^{M_{gen}} w_{gen,i}\, g^K(\boldsymbol{s}, \boldsymbol{\mu}_{gen,i}, \boldsymbol{\Sigma}_{gen,i}) \quad \textbf{(3-25)}$$

$$\hat{f}_{imp}(\boldsymbol{s}) = \sum_{i=1}^{M_{imp}} w_{imp,i}\, g^K(\boldsymbol{s}, \boldsymbol{\mu}_{imp,i}, \boldsymbol{\Sigma}_{imp,i}) \quad \textbf{(3-26)}$$

where M_{gen} and M_{imp} are the number of mixture components used to model the density functions of genuine and impostor scores respectively; $g^K(\boldsymbol{s}, \boldsymbol{\mu}_{gen,i}, \boldsymbol{\Sigma}_{gen,i})$ and $g^K(\boldsymbol{s}, \boldsymbol{\mu}_{imp,i}, \boldsymbol{\Sigma}_{imp,i})$ are the K-variate Gaussian density functions with mean vectors $\boldsymbol{\mu}_{gen,i}$ and $\boldsymbol{\mu}_{imp,i}$ and covariance matrices $\boldsymbol{\Sigma}_{gen,i}$ and $\boldsymbol{\Sigma}_{imp,i}$ of the i^{th} mixture component respectively; $w_{gen,i}$ and $w_{imp,i}$ are the weights assigned to the i^{th} mixture component.

The number of Gaussian components was selected using the algorithm proposed in [79], which automatically estimates the number of components and its parameters using an EM algorithm and the Minimum Message Length (MML) criterion. This algorithm is also robust to initialization of parameter values (mean vectors and covariance matrices) and it can handle discrete components in the match score distribution by modeling the discrete scores as a mixture component with very small variance.

3.4.1.3. Classifier-Based Score Fusion

Finally, this last approach for combining matcher's scores is based on the idea of using classifiers for finding the decision boundary between the classes, specifically, for the verification task, between genuine and impostor classes. In this

methodology, every score vector is treated as a feature vector in a classification scheme.

Because the classifiers treat the scores as features, it is irrelevant the form of the scores, i.e. they can be non-homogeneous, have different scales, etc. In this work, the probabilistic versions of two popular classifiers were used: 1) the Support Vector Machine (SVM) and 2) Random Forest (RF). These versions of algorithms are called probabilistic because they provide the posterior probability of each class for every input test pattern classified. In the following lines, they are succinctly explained.

Support Vector Machines

Support vector machines (SVM) are a set of supervised learning algorithms, introduced by Vapnik et al. in 1995 [80], used for classification and regression problems. It was firstly introduced for binary classification, but eventually it was extended to be used in multiclass classification too. It is based on the idea of seeking a decision boundary between two classes that maximizes the margin between them, and therefore minimizes the classification error.

Formally, given N training samples of dimension D, $\mathbf{x_i}, i = 1, \dots, N$ with $\mathbf{x} \in \mathbb{R}^{\boldsymbol{D}}$ that pertain to one of two classes $y_i \in \{-1,1\}^N$, it is desired to separate the classes with the hyperplane:

$$\mathbf{w} * \mathbf{x} + b = 0 \qquad \textbf{(3-27)}$$

where $\mathbf{w}$ is the vector normal to the hyperplane and b is the bias.

If the data is linearly separable, there exist an infinite number of planes that divide the classes, but obtaining the one that maximizes the margin can be shown to be equivalent to a minimization problem of the form:

$$\arg\min_{\mathbf{w},b} \frac{\mathbf{1}}{\mathbf{2}} \|\mathbf{w}\|^{\mathbf{2}} \quad \text{subject to} \quad y_i(\mathbf{w}^{\boldsymbol{T}}\mathbf{x}_i + b) \geq 1 \quad \forall_i \qquad \textbf{(3-28)}$$

If the data is not linearly separable, it is needed a transformation onto a high-dimensional feature space in which the data is linearly separable, by using a non-linear mapping function $\phi(\cdot)$. Since only the inner product of two vectors in that new space matters, the problem boils down to finding a function, called kernel function, that computes the inner product in that space, with no need to explicitly mapping from the low to the high dimensional space. Typical used kernels are the

Polynomial kernel, *Sigmoidal kernel* and *Radial Basis Function (RBF)*, the latter used in this work, with the form:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2) \quad \text{(3-29)}$$

where γ is a modulating constant.

When it is allowed some points to violate the margin, opening the possibility for some errors to occur, the margin is called soft margin. Considering this case, and using a kernel in a non-linearly separable problem, the minimization equation (3-28) can be extended in a general form as follows:

$$\begin{aligned} &\arg\min_{\mathbf{w},b,\xi} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \\ &\text{subject to} \quad y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \ \ \forall_i, \ \xi_i \geq 0 \end{aligned} \quad \text{(3-30)}$$

where $C > 0$ is a penalty parameter of the error ξ.

The solution to this problem is given after finding the Lagrangian multipliers $\boldsymbol{\alpha}$ in the equation:

$$\begin{aligned} &\arg\min_{\boldsymbol{\alpha}} \ \frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{\mathcal{H}}\,\boldsymbol{\alpha} - \sum_{i=1}^{N}\alpha_i \\ &\text{subject to} \quad 0 \leq \alpha_i \leq C, \ \forall_i \text{ and } \sum_{i=1}^{N}\alpha_i\, y_i = 0 \end{aligned} \quad \text{(3-31)}$$

where $\boldsymbol{\mathcal{H}}$ is the matrix formed with the elements $H_{ij} = y_i y_j \phi(\mathbf{x}_i)\phi(\mathbf{x}_j)$.

Using a Quadratic Programming optimization tool, the $\boldsymbol{\mathcal{H}}$ matrix is passed as input, and the $\boldsymbol{\alpha}$ values are returned. The values of $\boldsymbol{\alpha}$ different to zero are called the support vectors. With these values, $\mathbf{w}$ and b are computed following this procedure:

- compute $\mathbf{w} = \sum_{i=1}^{N}\alpha_i\, y_i\, \phi(\boldsymbol{x}_i)$
- find the indices of α_i such as $0 \leq \alpha_i \leq C$ (the support vectors) (3-32)
- solve for b using any $\alpha > 0$: $b = y_m - \sum_{\alpha_n>0}\alpha_n\, y_n\, k(\mathbf{x}_n, \mathbf{x}_m)$

Each new point $\mathbf{x}'$ is finally classified by evaluating $y' = sgn(\mathbf{w}\,\mathbf{x}' + b)$. The parameters C and γ should be estimated using cross validation in the training data.

Random Forests

The random forest algorithm was proposed by L. Breiman in 2001 [53], and it has become a very popular method for general-purpose classification and regression because of its simplicity to train and tune. In essence, it is an ensemble

of randomly trained decision trees, where each tree is constructed using a random subset of the data, and the features in the data, and then their results are averaged. This principle is called bagging, and it is the essential idea behind random forest. By averaging many noisy models, such as decision trees, the intrinsic variance of the trees is reduced, giving good generalization performance.

The algorithm pseudo code is as follows:

1. For $b = 1$ to B:
   (a) Draw a bootstrap samples $\mathbf{Z}$ of size N from the training data.
   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
      i. Select m variables at random from the p variables.
      ii. Pick the best variable/split-point among the m.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of the trees $\{T_b\}_1^B$.

When used for classification, each tree outputs a class prediction, and the decision is made using majority voting from all trees. When posterior probabilities are required, they are given by the ratio between the number of trees that selected a given class and the total number of trees in the ensemble. In contrast, if the algorithm is used for regression, the predictions from each tree at a target point x are averaged. A recommended values for m and for the minimum node size n_{min} are $\sqrt{p}$ and 1 respectively [81].

4
UNIMODAL BIOMETRIC SYSTEMS SETUP

In this chapter it is detailed the experimental setup designed to seek the optimal parameters for the unimodal biometric systems, and their performance is evaluated using two different datasets, which are also described in the course of this section. In addition, are presented the metrics used in the evaluation, the outlines of the experimental procedures, the protocols for configuring the unimodal recognition systems and the results of the experiments conducted for each configuration.

4.1.
Datasets

In this work, two bi-modal datasets comprising voices and faces were used to assess the performance of the unimodal systems. The first was artificially constructed from two individual unimodal datasets, specifically the TIMIT speech corpus [82] and the Facial Recognition Technology (FERET) dataset [83] (**Virtual Database** hereinafter). The other dataset used was the MOBIO database [18], which is a truly bi-modal dataset (**MOBIO Database** hereinafter).

It is worth remarking that the parameter tuning methodology that is presented in this chapter was conducted using only the Virtual database, this is, the composite of TIMIT and FERET databases, whereas the performance evaluation was completed for both databases. For the MOBIO database, the parameters were fixed using the best configuration obtained from the experimental analysis of the Virtual database.

Virtual Database

The virtual bimodal database was constructed using the TIMIT and FERET databases which are going to be explained in detail in the course of this section.

The Speaker Recognition System (SRS hereinafter) was tuned using the TIMIT corpus. The corpus consists of native American English speakers from 8 dialect regions. It contains in total 630 speakers, of which 438 are males (70 %) and 192 females (30 %), according to the distribution shown in the table below (Table 4-1).

Dialect Region	Males	Females	Total
1	31 (63%)	18 (27%)	49 (8%)
2	71 (70%)	31 (30%)	102 (16%)
3	79 (67%)	23 (23%)	102 (16%)
4	69 (69%)	31 (31%)	100 (16%)
5	62 (63%)	36 (37%)	98 (16%)
6	30 (65%)	16 (35%)	46 (7%)
7	74 (74%)	26 (26%)	100 (16%)
8	22 (67%)	11 (33%)	33 (5%)
Total	438 (70%)	192 (30%)	630 (100%)

Table 4-1: Dialect distribution of speakers in TIMIT database.

There are 10 speech files for each speaker (hereinafter, samples per person). Two of the files have the same linguistic content for all speakers, whereas the remaining 8 files are phonetically diverse. The corpus has been recorded in a soundproof environment with a high-quality microphone. Speech files are stored in NIST/Sphere "wav"-file format with a sampling frequency of 16 kHz and a quantization resolution of 16 bits per sample.

In order to reproduce a real operation scenario, the utterances were mixed with artificial noise at different Signal to Noise Ratios (SNR). For this purpose, in addition to the TIMIT corpus, noise samples from the Noisex-92 database were used [84-86]. Specifically, white noise, pink noise, speech babble, factory noise, car noise and f16 noises were chosen to mix with the speaker voices. The noises were randomly added to the speech data, before the feature extraction, as described in the Section 4.4.1.

For the configuration of the Facial Recognition System (FRS hereinafter), it was used the FERET (Facial Recognition Technology) dataset [83]. It contains a total of 14,126 images that includes 1199 individuals.

Images of individuals are in sets of 5 to 11 images, so different quantity of images are available from some persons. Common to all sets, there are two frontal views, *fa* and *fb*, the first being a neutral pose while the second was taken using a different facial expression. The rest of images in the sets have variations in scale, pose, facial expression, illumination, facial accessories (i.e. glasses) and rotation, as can be seen in the sample below (see Figure 4-1) [87].

Figure 4-1: Sample of Images from FERET database.

The facial images were geometrically and photometrically normalized using the coordinates of the eyes, provided as metadata in the FERET database. Not every person in the database had that information, so only 866 individuals of the total were normalized, ending with 366 females and 500 males after normalization. Also the number of normalized frontal images per person varies, depending on the set of images.

The construction of this combined database was accomplished by creating virtual persons (or chimeras[1]), randomly pairing a user from one unimodal database with a user from the other database. Because the FERET dataset is larger than the TIMIT, some users were randomly discarded, ending up with the same number of individuals and the same proportion of males and females, i.e., 192 females and 438 males.

The composite of data between the datasets implicitly assumes that the two biometrics (face and speech) are independent, which has been studied to be a valid approximation, from prior related works. It was considered the gender in the pairing of both databases as well.

[1] Chimeras are composites of data representing virtual "individuals" that combine biometrics from multiple individuals (selected at random).

MOBIO Database

The MOBIO database is a truly bi-modal database of audio and video captured by mobile phones and laptops. It comprises 150 individuals, of which 99 are males and the rest females. The data was recorded in 6 different locations in 5 different countries, and in 12 different sessions, with people speaking English. Samples of this database are shown in Figure 4-2. This database is challenging because the images and utterances were extracted under adverse conditions, i.e. uncontrolled illumination, background noise, facial expressions, occlusion, and other effects, as it can be seen in the figure.

Figure 4-2: Samples of Images from MOBIO database. It shows two individual under different session conditions, where occlusion, illumination and pose effects are present.

For this dataset, the International Conference on Biometrics (ICB-2013) opened a competition in which two evaluation protocols for speaker and face recognition systems were defined [88, 89]. In those, an identical partitioning of the database was set on each case, as shown in Table 4-2.

	Training		**Development**				**Evaluation**			
			Enrollment		**Probe**		**Enrollment**		**Probe**	
	Clients	**Files**	**Clients**	**Files**	**Files**	**Scores**	**Clients**	**Files**	**Files**	**Scores**
male	37	7104	24	120	2520	60480	38	190	3990	151620
female	13	2496	18	90	1890	34020	20	100	2100	42000
Total	50	9600	42	210	4410	94500	58	290	6090	193620

Table 4-2: Partitioning of the MOBIO database in Training, Development and Evaluation sets for the ICB-2013 evaluation competition.

In this work, we used the same partitioning for our experiments, with the difference that we discarded the evaluation set because the probe samples were not labeled, so we only worked with the training and development sets. Therefore, the total number of persons we used was 92.

4.2. Metrics

For assessing the performance of the biometric classifiers in identification and verification modes, we used the CMC (*Cumulative Match Characteristic*) curves and the ROC (*Receiver Operating Characteristic*) curves respectively.

The CMC curve describes the proportion of times in which the correct classification of probe samples is observed within the top k ranks. In other words, a rank-1 outcome for a given probe is considered a correct identification and a rank-5 result means the correct identity is within the top 5 ranks of the score set. The ranking process is repeated for every probe individual and at the end, the percentage of correct matches is computed for each rank value.

The ROC curve on the other hand describes the behavior of classifiers operating in verification mode, based on the metrics of *False Acceptance Rate (FAR)* or *False Positive Rate (FPR)* and *False Rejection Rate (FRR)* or *False Negative Rate (FNR)*. Two types of errors can occur: (1) *false rejection*, that is, falsely rejecting a genuine user's claim, or (2) *false acceptance*, that is, falsely accepting the claim to be from a genuine user when the actual person is an impostor. These curves are constructed varying a threshold t at which an individual is accepted or rejected. The error rates can be computed as follows:

$$FRR(t) = FNR(t) = \frac{\left|\{s_{gen} | s_{gen} < t\}\right|}{\left|\{s_{gen}\}\right|} \quad \textbf{(4-1)}$$

$$FAR(t) = FPR(t) = \frac{\left|\{s_{imp} | s_{imp} \geq t\}\right|}{\left|\{s_{imp}\}\right|} \quad \textbf{(4-2)}$$

The FRR represents the percentage of genuine scores that are below the threshold and are incorrectly classified as impostors whereas the FAR denotes the percentage of impostor scores that exceed the threshold and are incorrectly classified as genuines. The term s_{gen} refers to the genuine scores and s_{imp} refers to the impostors.

Other metrics used throughout this work are the *Genuine Acceptance Rate* or *True Positive Rate (GAR = TPR = 1-FRR),* the *equal error rate* (*EER*), which is the point where FAR equals FRR, and it is a measure of the authentication accuracy at the decision threshold of the classification system, the *Area under ROC Curve* (*AUC*), which is a measure of discrimination between genuine from impostors.

4.3. Outline of the Experimental Setup

Since there is a large number of adjustable parameters to be evaluated, the assessment of all possible parameter combinations is not possible. Therefore it is applied a simple *line search* strategy, in which one parameter is varied at a time while keeping the rest fixed. Although this procedure does not guarantee a globally optimal parameter combination, it gives an idea what are the most critical parameters that need to be adjusted in a real application scenario.

4.4. Unimodal Biometrics Evaluation

In this section, they are described the experiments designed to find the optimal parameters for both SRS and FRS.

4.4.1. Speaker Recognition System Evaluation Protocol

For the SRS, it was used the Mel-Frequency Cepstral Coefficients (MFCC) as the voice features in the front-end and the Gaussian Mixture Model with Universal Background Model (GMM/UBM) and the I-Vector Model with Gaussian Probabilistic Linear Discriminant Analysis (GPLDA) as the back-end. The parameters of the front-end (MFCC) were fixed in order to provide the state-of-the-art performance in this application, according to the literature [63]. In this case, it was configured as detailed in Table 4-3. Thus, the voice features has dimension 39 (13 Cepstral Coeff. + 13 Delta features + 13 Delta-Delta features). It is important to point out that the length of the utterances are variable, so the matrix that describe the voice features have a variable size in one of its dimensions.

Frame Duration = 25mseg	*# Filter Bank Channels = 30*
Frame Shift = 10mseg	*# Cepstral Coeff. = 13*
Preemphasis Coeff. (α) = 0.97	*Energy Value = Yes*
Windowing Function = Hamming	*Delta & Delta-Delta Coeff. = Yes*

Table 4-3: MFCC Parameters for Speaker Recognition System

For the I-Vector Model, a total variability space T with 400 total factors was learned from the GMM models supervector, which is formed by concatenating the mean components of the GMM models into a high and fixed dimensional single vector, (see Section 3.2.2). This dimension of T has been found to be good for the speaker recognition problem, according to the literature [69].

For the back-end, the performance of the classifiers was evaluated varying the number of samples per person in Gallery and the number of Gaussian components to be used in the GMM/UBM and the GPLDA. Also, it was important to measure the performance of the system with the addition of noise in the data by changing the Signal to Noise Ratio (SNR).

To start with all possible combinations in the system performance evaluation, it was created a base configuration with the following parameters:

- *No Noise*
- *1 Sample per Person in Gallery*
- *128 Gaussian Components*
- *80% Training Set and 20% Test Set*

The training and test sets represent the percentage of all speakers that were used for creating the Universal Background Model (UBM) and for testing the system respectively. The gender information was used to balance the training step, so the same percentage of females and males were used, and not to create two UBMs, as it is reported in other related works (e.g., [65]). In addition, every utterance of each training speaker (10 utterances per speaker in TIMIT) was used for training the UBM. The reason for this was to pool the maximum amount of speech data to create the UBM, in an effort to universally represent the person-independent feature characteristics. For gallery enrollment, the persons not used in the training process were used with the configured number of samples per person in gallery in the adaptation stage. In the test step, only one sample per person was used, different from those used in gallery enrollment.

Using the base configuration described above, three experiments were defined and are presented in Table 4-4. Each experiment took in consideration the best result from the previous one.

For every experiment, the dataset was divided in 5 equally populated and disjoint groups (i.e., folds). It was used 80% of the database to train the UBM (4

folds) and 20% to test the performance of the system (1 fold). The experiments were run 5 times, choosing a different test set each time, until 5 different combinations were covered. Finally, the resulting CMC and ROC curves were averaged among all groups, obtaining cross-validating results.

Experiment Number	Target Parameter	Procedure
1	Noise	Using the base configuration, test the system performance using No Noise and different levels of SNRs (24dB, 12dB, 0dB respectively).
2	Number of samples per person in gallery	Using the base configuration, test the system performance using 1, 5 and 9 samples per individual in gallery.
3	Number of Gaussian Components	Using the base configuration, test the system performance using 32, 64, 128, 256 and 512 Gaussian Components.

Table 4-4: Experimental Configuration for Speaker Recognition System Evaluation

4.4.2. Results

For the **Experiment 1**, Figure 4-3 and Figure 4-4 show the CMC and ROC curves of the SRS under different noise conditions.

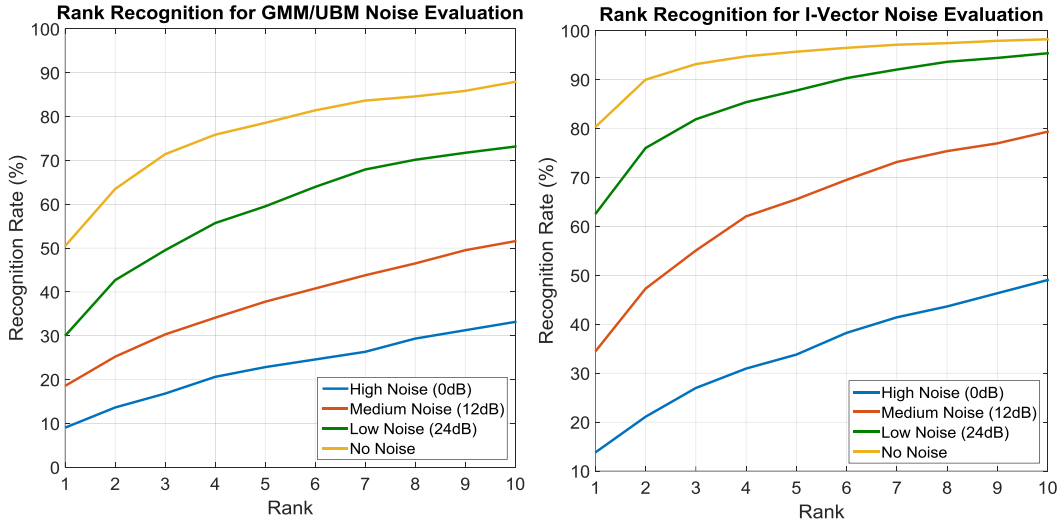

Figure 4-3: CMC curves for Noise Evaluation in Speaker Recognition System using GMM/UBM and I-Vector techniques.

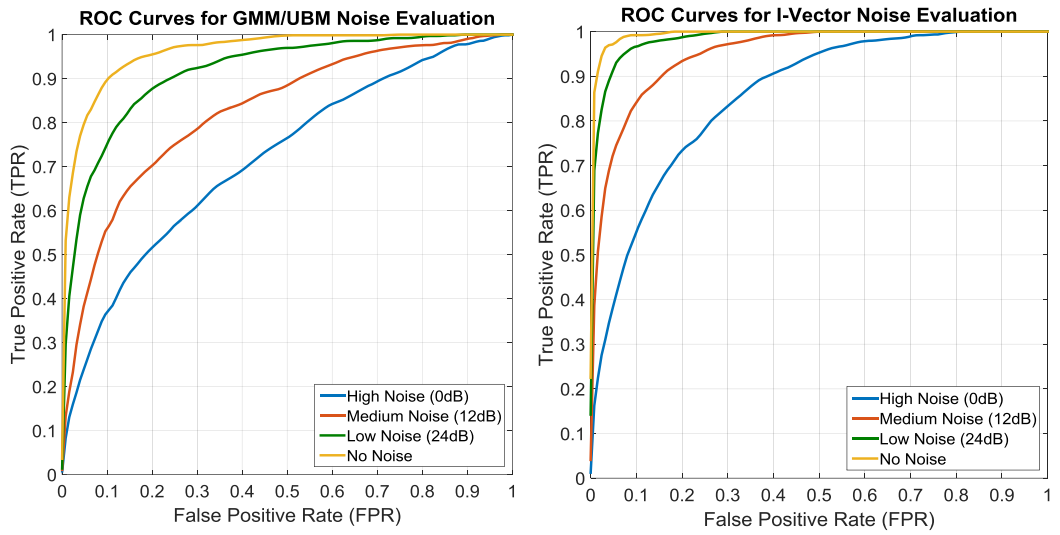

Figure 4-4: ROC curves for Noise Evaluation in Speaker Recognition System using GMM/UBM and I-Vector techniques.

It can be observed how the recognition accuracy degrades with higher SNRs, with the I-Vector framework offering the most robust behavior of both techniques under noise.

The results of **Experiment 2** can be seen in Figure 4-5 and Figure 4-6.

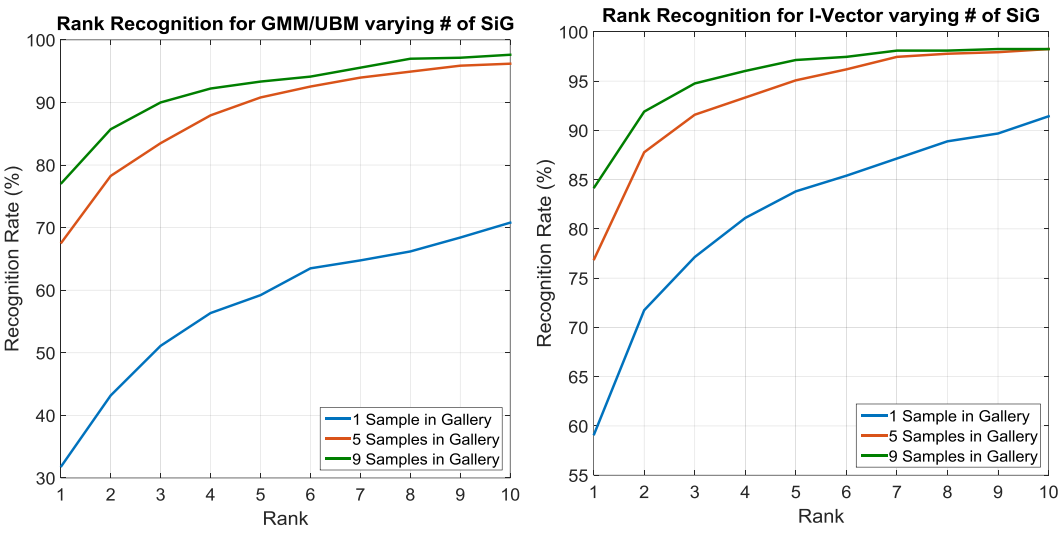

Figure 4-5: CMC curves for different number of samples per person in gallery (SiG) for GMM/UBM and I-Vector techniques.

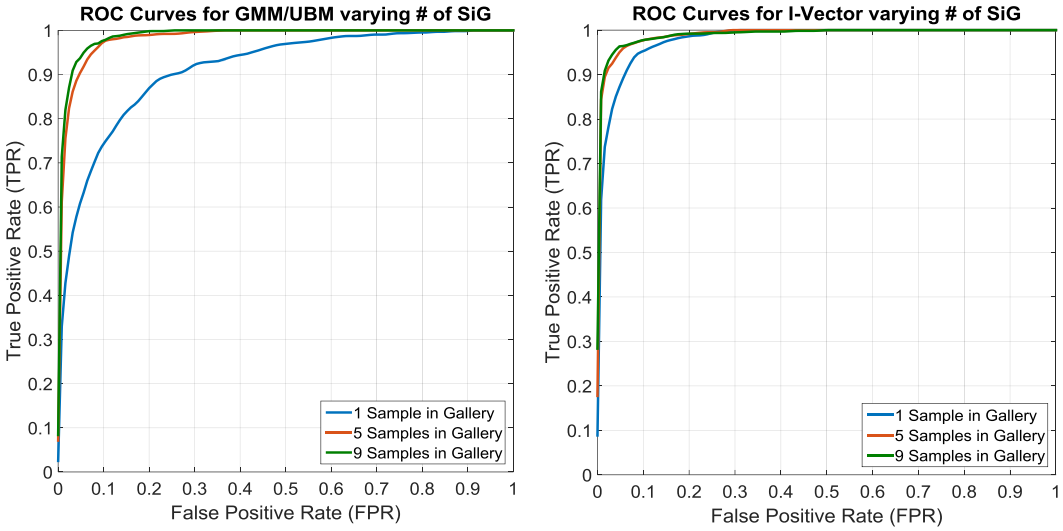

Figure 4-6: ROC curves for different number of samples per person in gallery (SiG) for GMM/UBM and I-Vector techniques.

As expected, as the number of samples for enrollment grows, the recognition accuracy increases, reaching high Recognition Rates, near 85% in the I-Vector approach. The reason for this behavior is that using more data of every user, the system can create better models.

Figure 4-7 and Figure 4-8 show the results for the last experiment (**Experiment 3**), varying the number of Gaussian Components in the models.

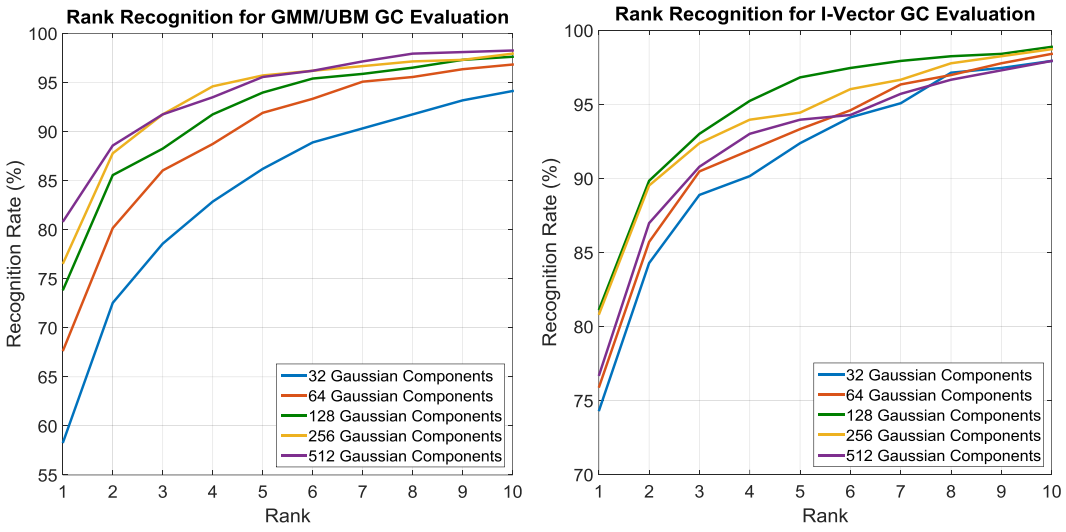

Figure 4-7: CMC curves for different number of Gaussian Components (GC) for GMM/UBM and I-Vector techniques.

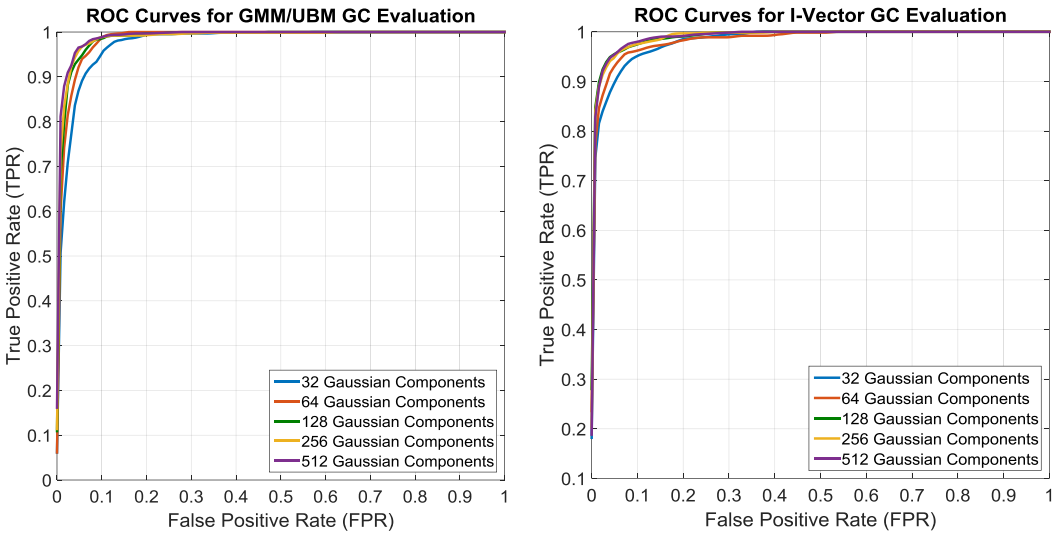

Figure 4-8: ROC curves for different number of Gaussian Components (GC) for GMM/UBM and I-Vector techniques.

For GMM/UBM approach, the best performance was obtained with the highest Gaussian components, i.e. 512, whereas for the I-Vector framework, 256 Gaussians provided the best result.

4.4.3. Facial Recognition System Evaluation Protocol

For the FRS, two different feature extraction techniques were implemented: (1) the DCT-based feature extractor seen in Section 3.3.1 for the GMM/UBM classifier and (2) the LBP operator with the square-chi distance classifier (hereinafter *LBP-based classifier*).

In order to find which algorithm provides the best recognition performance, it was also necessary to adjust the parameters for the GMM/UBM classifier, in a similar manner as with SRS. The feature extraction parameters for the GMM/UBM and for the LBP-based classifier were fixed, according to reference values taken from other related works [76, 90, 91], and they are shown in the Table 4-5 below:

Features for GMM/UBM	Features for LBP-based classifier
Block Size = 16 x 16 pixels	Number of equally spaced pixels = 8
Overlap = 50%	Radius = 2
Illumination Compensation = Yes	Block Size = 8 x 8 pixels
Dimension of the feature vectors = 63	Number of Bins in Histogram = 59

Table 4-5: Features used for GMM/UBM and LBP-based classifiers.

The parameters of the GMM/UBM classifier used in the FRS were tuned in the same way as in SRS. In this case, the performance was measured with different number of Gaussian components using 80% of the total amount of users as training set and 20% as test set.

It is important to point out that due to the arrangement of the FERET dataset where only few individuals have a duplicate set of images, a considerable amount of persons in the database had only a pair of useful frontal images available for training and testing, after the database normalization. For this reason, it was used only one image per person for gallery adaptation and only one image for testing, specifically the FERET database's *fa* image for gallery and *fb* for testing respectively. However in the training process of the UBM, all available images were used.

Using the aforementioned configuration, the experiment presented in Table 4-6 was defined.

Experiment Number	Target Parameter	Procedure
4	Number of Gaussian Components	Test the system performance using 32, 64, 128, 256 and 512 Gaussian Components and compare with LBP approach.

Table 4-6: Experimental Configuration for Facial Recognition System Evaluation

Like the SRS experiments, an analogous procedure was followed for the FRS evaluation, dividing the dataset in 5 equally populated and disjoint groups: 4 groups used for training the UBM and 1 group for testing. The experiments were run 5 times, choosing a different test set each time, covering 5 different combinations. The resulting CMC and ROC curves were averaged among all groups, as in the SRS. Finally, the best GMM/UBM configuration was compared with the LBP-based classifier.

4.4.4. Results

The results of experiment 4 are shown in Figure 4-9 and Figure 4-10. As it can be noted, the LBP-based approach provided the best rank recognition accuracy, managing to reach almost 90% in rank-1 recognition rate. The best configuration

of GMM/UBM was obtained using 512 Gaussian Components, for a rank-1 recognition rate of near 83%. The ROC curves showed similar performance between the two approaches.

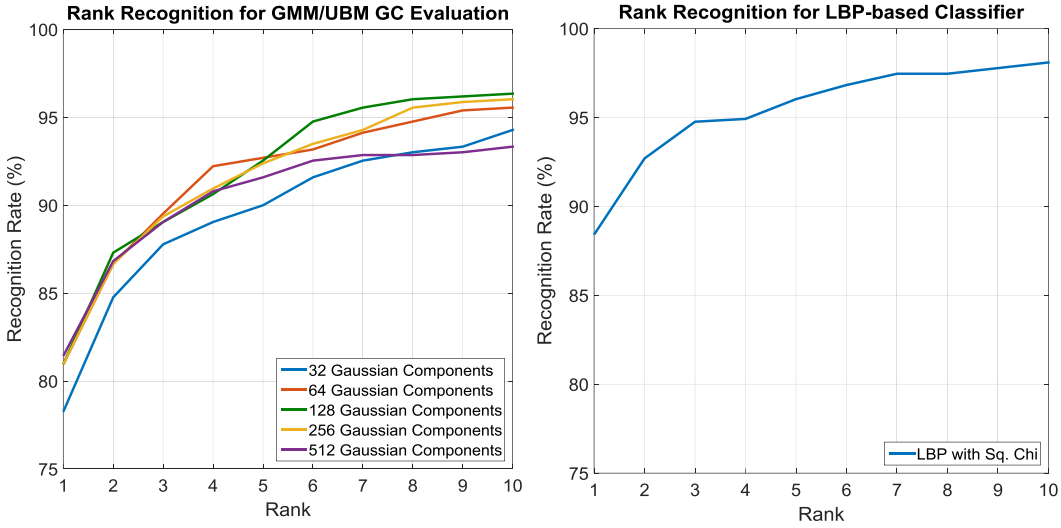

Figure 4-9: CMC curves comparing GMM/UBM with different Gaussian Components and LBP-based Classifier.

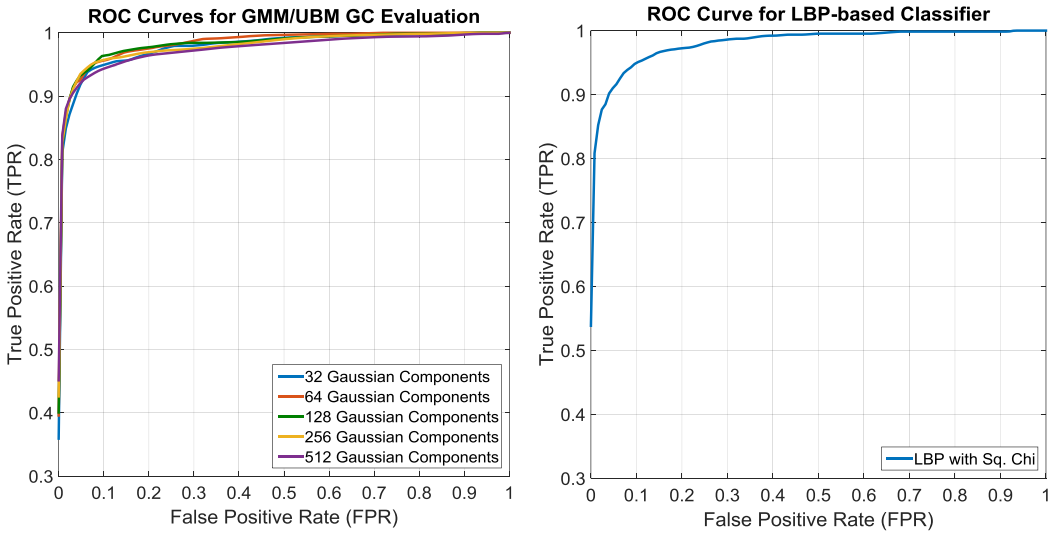

Figure 4-10: ROC curves comparing GMM/UBM with different Gaussian Components and LBP-based Classifier.

4.5. MOBIO Evaluation

To evaluate the performance of the unimodal systems using the MOBIO dataset, we selected the same feature extraction parameters chosen for the other database. The selection of the training and test sets followed the protocols defined in ICB-2013 for face and voice [92], where the number of samples per individual

for gallery enrollment was identical for both face and speaker protocols, being 5 samples per individual, and 105 the number of samples used as probe. The entire training set defined in the aforementioned protocol was used to create the UBMs in the SRS and FRS systems and was also used to train the GPLDA hyperparameters in the I-Vector algorithm. For the LBP scheme, the scores generated by each of the five gallery samples were averaged for each algorithm.

One modification in our procedure with respect to the ICB protocol is that we tested each probe file with all the persons in the gallery, and we made no distinction between females and males. This generated 4410*42 (185220) scores instead of 94500 that is reported in the original protocol.

4.5.1. Results

SRS Evaluation

Figure 4-11 Figure 4-12 show the Rank Recognition and the ROC curves respectively of the SRS when tested using the MOBIO database. With this database, the SRS obtained a Rank-1 recognition rate lower than that of the other database. As an example, in the results of the **Experiment 3**, the SRS system reached a Rank-1 recognition rate of almost 82% in both GMM/UBM and I-Vector approaches, whereas for this database it managed to obtain only 66% of recognition accuracy for the same algorithms. This difference in the results were expected, because the MOBIO audio is far noisier than the TIMIT corpus used in the first database, and also because it was acquired using different microphones and locations, whereas the TIMIT data was recorded in a controlled environment with one high-quality microphone.

Between the two algorithms, we can see that the I-Vector provides the best result for both identification and verification tasks.

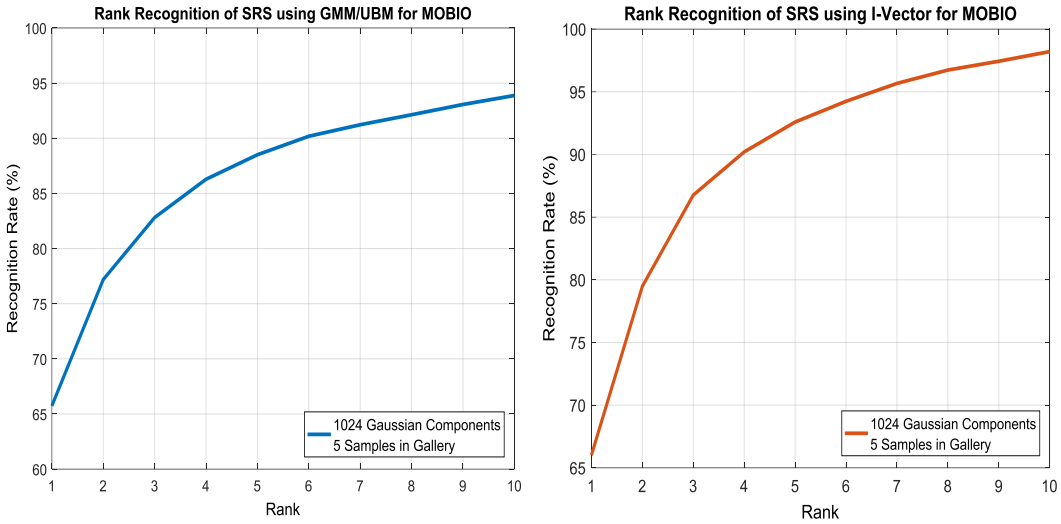

Figure 4-11: CMC curves comparing GMM/UBM and I-Vector approaches for the SRS using the MOBIO dataset.

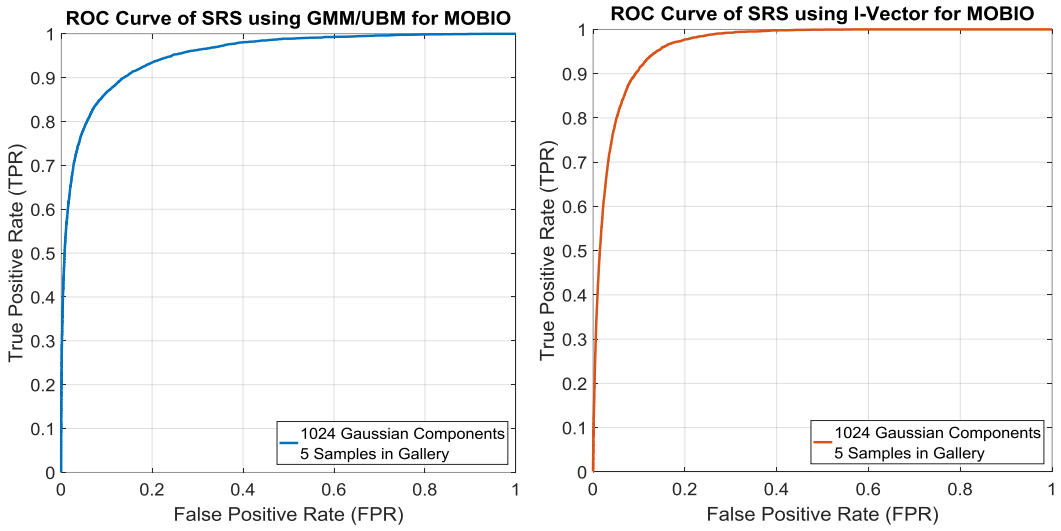

Figure 4-12: ROC curves comparing GMM/UBM and I-Vector approaches for the SRS using the MOBIO dataset.

FSR Evaluation

In Figure 4-13 and Figure 4-14 are shown the results of the Rank Recognition rates and ROC curves for the FRS using MOBIO database. It can be noted, similar to the SRS evaluation, that the results are not great for either algorithm tested, as the images in this database contain several nuisance effects, such as occlusions, illumination problems, pose variations and others. Between GMM/UBM and LBP-based Classifier approaches, the former obtained better results than the latter. The LBP obtained low recognition rates because it relies heavily on the frontal pose of the facial images, and any significant variation may affect dramatically the accuracy

of this algorithm. In contrast, the GMM/UBM, as it is a probabilistic approach, obtained better results, as it managed to model more efficiently these variations.

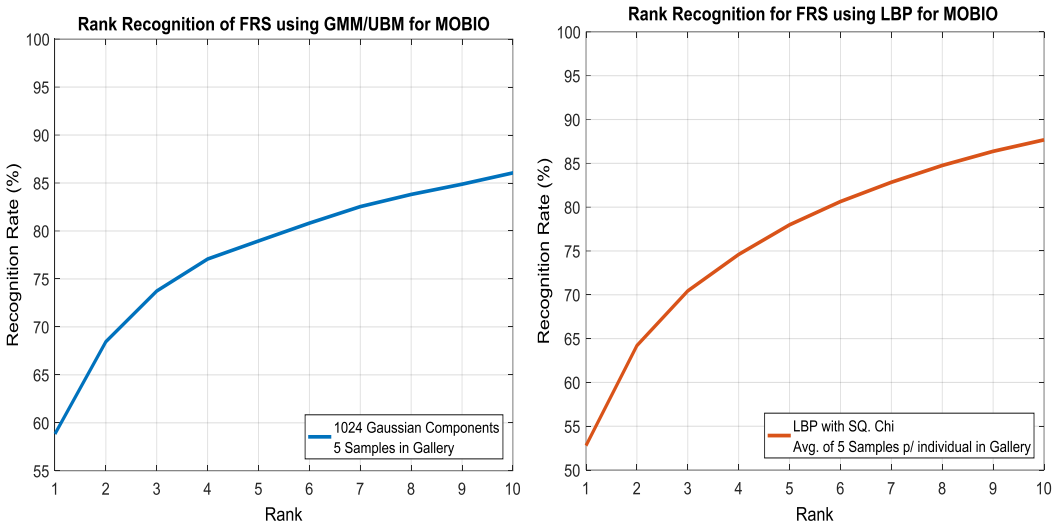

Figure 4-13: CMC curves comparing GMM/UBM and LBP-based Classifier approaches for the FRS using the MOBIO dataset.

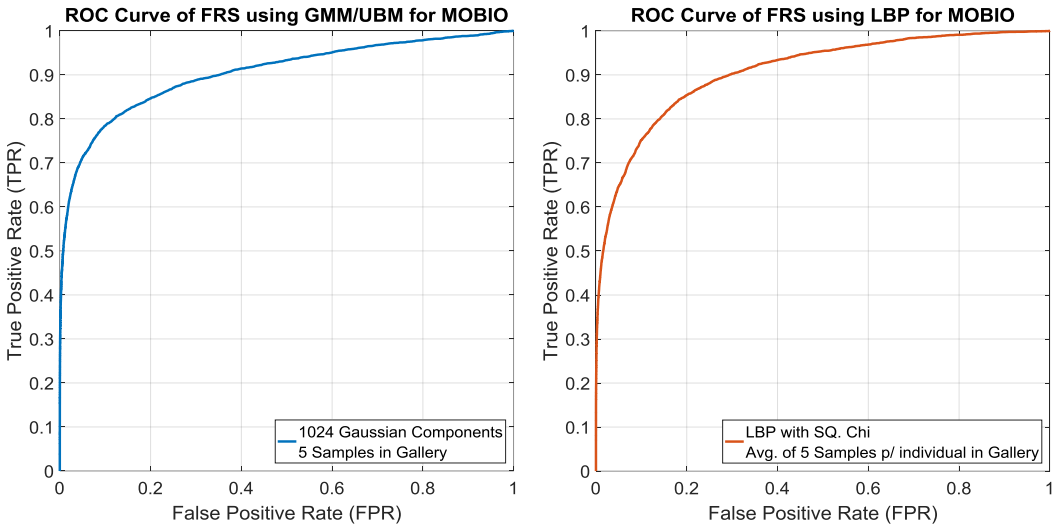

Figure 4-14: ROC curves comparing GMM/UBM and LBP-based Classifier approaches for the FRS using the MOBIO dataset.

4.6. Summary

In this chapter, we followed an experimental methodology for configuring the unimodal biometric systems that are to be fused. For this purpose, we tested the algorithms developed using two very different databases, one formed by two independent unimodal databases, and the other a truly bi-modal database.

The results obtained here confirmed that when the unimodal biometrics are restricted to application environments with high variations and disturbance effects, such as noise, pose variations, occlusions and location changes, the recognition and authentication performance of these systems degrades considerably. This motivated the use of multimodal biometric fusion, described in the next chapter, in an effort to overcome these difficulties.

5
MULTIMODAL FUSION

This chapter describes the experimental framework developed for comparing the fusion schemes proposed in this dissertation, applied to the fusion of the face and speaker recognition systems presented. It is described the methodology used, the implementation details of techniques and algorithms used and finally they are presented the results of the experiments.

5.1.
Multimodal Training/Test Sets Setup

The experiments conducted in this chapter are based on the scores generated by the speaker and facial unimodal biometric systems, using the databases described in Chapter 4, specifically the Virtual database and the MOBIO database. Two different groups of experiments are presented, one for each database.

For the MOBIO database, the scores for the fusion were generated using the same dataset partitioning described in the previous chapter, i.e., 50 users for training the models and 42 for gallery enrollment and test. A total of 110 samples for each of these 42 individuals are available, and we used 5 samples for gallery enrollment and 105 samples for test. Since 42 individuals were enrolled in gallery, the probes samples added up to 42*105 (4410) in total. Each of these probe samples was compared with each model in the gallery, yielding 185220 scores, with 4410 genuines and the remaining impostors.

For the Virtual database, a different partition of the dataset was used to generate the scores for the fusion. In this case, the database was divided in two halves, one half used for training the individual classifiers and the other half to generate the scores. For generating the scores, each sample of every test user was matched with the rest of the identity models enrolled in gallery. In this work, with a database of 630 users, 50% of users were used for training and 50% for test, and one sample per test user was used, which yielded 315 genuine scores and 98910

impostor scores. For this dataset, this procedure was repeated two times, switching the training and test sets. Finally, the results of the fusion stage were averaged for each execution.

5.2. Multimodal Verification Evaluation

In this section, it is described the methodology used to combine the matcher scores from both unimodal systems. The experiments carried out evaluated the performance of the multimodal fusion in verification mode. Three main approaches were implemented: 1) Density-Based Score Fusion, 2) Transformation-Based Score Fusion and 3) Classifier-Based Score Fusion.

For every fusion approach tested, the scores were divided in five disjoint groups, where four groups (4 folds) were used for training the parameters of each fusion scheme, and the remaining group was used to test the performance of the system (1 fold). The experiments were run 5 times, choosing a different test set each time and averaging the metric used at each case (i.e. ROC curves, Classification Accuracy, EER, etc.).

In Table 5-1 are summarized the techniques implemented in this operational mode.

Approach	Main Technique	Fusion
Density-Based Score Fusion	Kernel Density Estimation (KDE)	Sum and Product of Marginal Densities and Likelihood Ratio
	Mixture of Gaussians (MoG) Joint density estimation	Likelihood Ratio
Classifier-Based Score Fusion	Support Vector Machine (SVM)	Joint Scores Classification
	Random Forest (RF)	
Transformation-Based Score Fusion	Minmax Normalization	Average Sum and Product of Normalized Scores
	Z-Score Normalization	
	Tanh Normalization	

Table 5-1: Score Fusion Scheme for Experimental Evaluation in Verification Mode

Having two classifiers implemented for speaker recognition and two for facial recognition, four different fusion combinations were explored: a) GMM/UBM

Voice and GMM/UBM Face; b) GMM/UBM Voice and LBP-based Face; c) I-Vector Voice and GMM/UBM Face; d) I-Vector Voice and LBP-based Face.

5.2.1. Transformation-based Fusion

The first approach tested was the Transformation-Based Score Fusion because it is the simplest one and also because there are several combinations available for normalization and fusion.

The scores generated by the LBP-based classifier represent a distance measure while the rest of the scores corresponding to the other algorithms have a similarity meaning. Therefore the LBP scores had to be converted to a similarity metric using a simple distance-to-similarity transformation, subtracting the highest score from the rest.

5.2.1.1. Virtual Database Results

In Figure 5-1 are presented the resulting ROC curves of the experiments using the Virtual database for the unimodal and multimodal systems using the transformation-based fusion approach. In all figures presented hereinafter, ***MM***, ***Z*** and ***Th*** stand for Minmax, *z*-score and *tanh* normalization schemes, whereas Sum and Prod represent the Sum and Product fusion schemes.

From the figure, it can be observed that using this database, a multimodal system employing the average sum and product of normalized scores provides in general a better performance than the best unimodal system for all normalization techniques except the product of *z*-score in the last combination. As an example, at a FAR of 0.1%, the best unimodal module is the LBP-based matcher with a Genuine Acceptance Rate (GAR) of about 70%, while the Sum of Minmax normalized scores in that configuration generated a GAR close to 86%. This improvement in performance is significant and it underscores the benefit of multimodal systems.

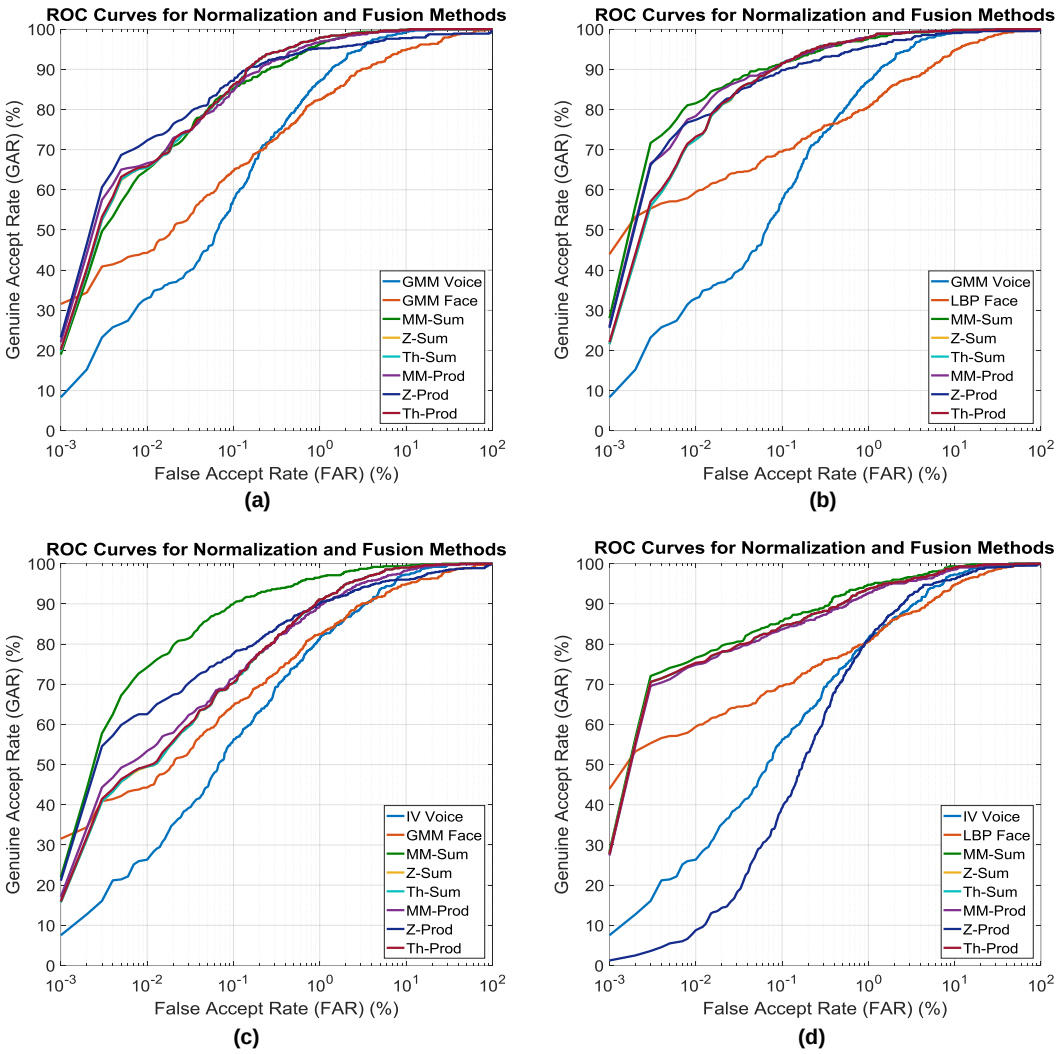

Figure 5-1: ROC Curves for Transformation-based Score Fusion in Virtual database. (a) Fusion of GMM/UBM Voice and GMM/UBM Face. (b) Fusion of GMM/UBM Voice and LBP Face. (c) Fusion of I-Vector Voice and GMM/UBM Face. (d) Fusion of I-Vector Voice and LBP Face.

In Table 5-2 are summarized the Equal Error Rate (EER), the Area under the ROC Curve (AUC) and GAR for this experiment. In yellow they are highlighted the techniques that provided the minimum EER for every combination of classifiers.

In general, the Minmax normalization technique provided the best results for the configurations b), c) and d), whereas the *tanh* normalization was better in configuration a). With respect to the fusion rules, the sum rule proved to get consistently lower EERs in all configurations than the product rule.

	Normalization Technique	Fusion Technique					
		Average Sum			Product		
		GAR[2]	AUC	EER	GAR[2]	AUC	EER
GMM Voice - GMM Face	**Minmax**	85,24	0,998	1,678	84,62	0,998	1,886
	Z-Score	86,05	0,999	1,515	87,14	0,986	3,499
	Tanh	86,05	0,999	1,514	86,21	0,999	**1,506**
GMM Voice - LBP Face	**Minmax**	90,02	0,998	1,476	71,52	0,998	**1,310**
	Z-Score	70,32	0,999	1,408	77,62	0,993	2,545
	Tanh	70,32	0,999	1,408	70,41	0,999	1,405
IV Voice - GMM Face	**Minmax**	91,59	0,998	**1,908**	91,36	0,992	4,139
	Z-Score	91,45	0,995	3,33	89,84	0,982	4,842
	Tanh	91,45	0,995	3,33	91,52	0,995	3,366
IV Voice - LBP Face	**Minmax**	85,71	0,996	**3,179**	83,65	0,994	4,075
	Z-Score	84,46	0,995	3,492	39,23	0,984	5,178
	Tanh	84,46	0,995	3,492	84,60	0,995	3,492

Table 5-2: Genuine Acceptance Rate (GAR), Area under ROC Curve (AUR) and Equal Error Rate (EER) of different normalization and fusion techniques for all classifiers combinations using the Virtual database.

5.2.1.2. MOBIO Database Results

Figure 5-2 shows the results of this fusion scheme for the MOBIO database. Similar to the previous experimental example, it can be noted the bi-modal fusion provided better results than the best unimodal system individually, for every configuration tested.

The higher gains in recognition accuracy were obtained in configurations a) and d), where the use of the facial biometric made possible to alleviate the inferior recognition rates offered by the I-Vector algorithm used in the speaker recognition system in this case.

[2] At FAR = 0.1%

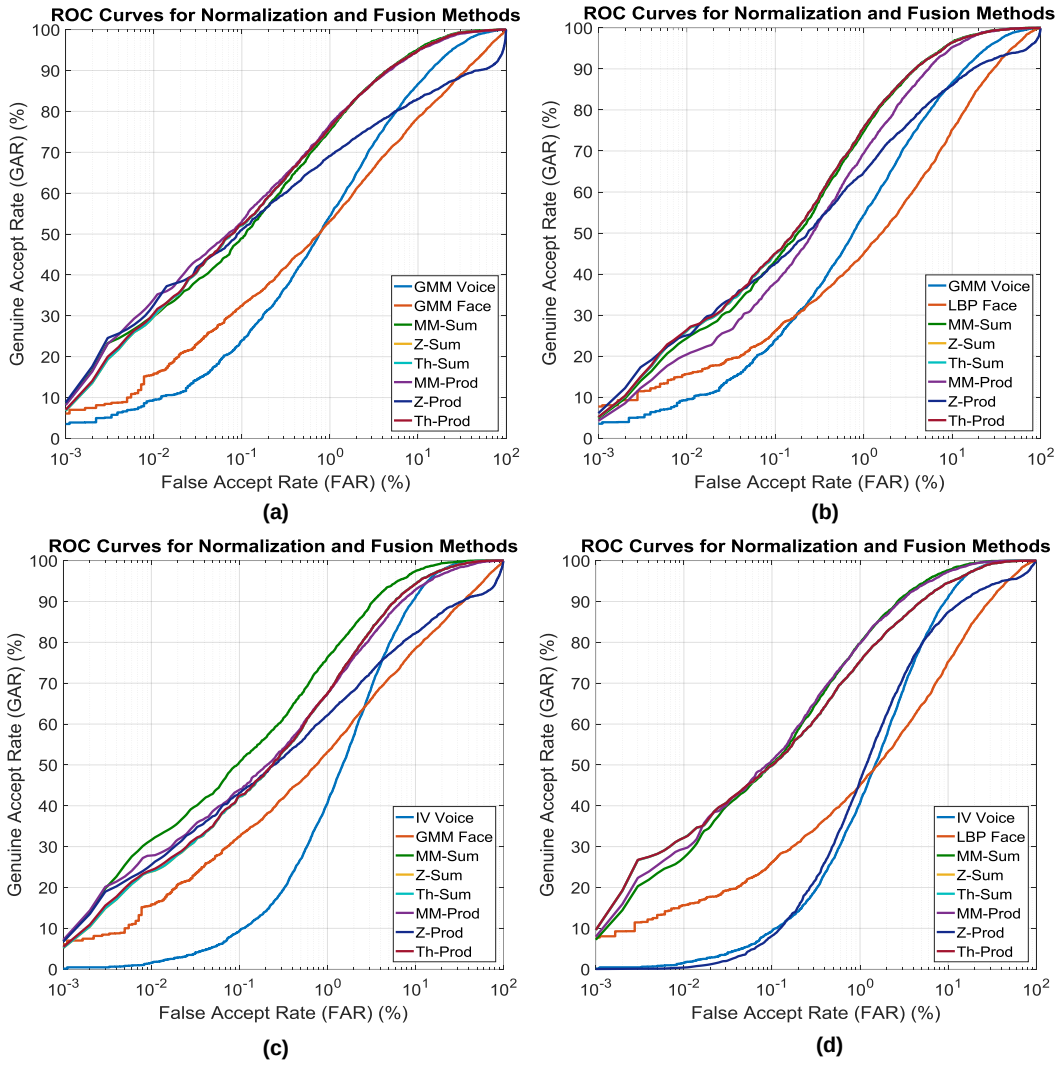

Figure 5-2: ROC Curves for Transformation-based Score Fusion in MOBIO database. (a) Fusion of GMM/UBM Voice and GMM/UBM Face. (b) Fusion of GMM/UBM Voice and LBP Face. (c) Fusion of I-Vector Voice and GMM/UBM Face. (d) Fusion of I-Vector Voice and LBP Face.

Comparing the normalization and fusion rules for this database, all of them showed similar results, except for the product of *z*-normalized scores, which showed poor performance. A more detailed view of the results is summarized in Table 5-3. Here the Average Sum rule with the Minmax normalization scheme dominated in every configuration tested, in relation to the EER.

	Normalization Technique	Fusion Technique					
		Average Sum			Product		
		GAR	AUC	EER	GAR	AUC	EER
GMM Voice - GMM Face	**Minmax**	48,98	0,982	**6,984**	53,01	0,980	7,193
	Z-Score	52,10	0,981	7,073	51,03	0,890	15,00
	Tanh	52,12	0,981	7,073	52,33	0,981	7,095
GMM Voice - LBP Face	**Minmax**	43,19	0,984	**6,361**	37,94	0,980	7,188
	Z-Score	44,95	0,984	6,433	42,57	0,923	12,29
	Tanh	44,98	0,984	6,433	45,06	0,984	6,434
IV Voice - GMM Face	**Minmax**	50,64	0,987	**5,639**	43,90	0,972	8,503
	Z-Score	42,20	0,978	7,570	43,19	0,900	14,86
	Tanh	42,21	0,978	7,571	42,44	0,977	7,585
IV Voice - LBP Face	**Minmax**	50,34	0,989	**5,319**	51,15	0,988	5,599
	Z-Score	49,76	0,980	7,237	8,15	0,930	11,59
	Tanh	49,76	0,980	7,237	49,89	0,980	7,236

Table 5-3: Genuine Acceptance Rate (GAR), Area under ROC Curve (AUR) and Equal Error Rate (EER) of different normalization and fusion techniques for all classifiers combinations using the MOBIO database.

5.2.2. Density-based Fusion

Next, it was evaluated the Density-Based Score Fusion. These methods are based on the Likelihood Ratio and the Neyman-Pearson rule described in Section 3.4.1.2. The standard deviation for computing the kernel bandwidth in KDE technique was estimated from the training scores, as well as the number of Gaussians in the MoG scheme, using the Minimum Message Length criterion described in [79].

5.2.2.1. Virtual Database Results

Figure 5-3 shows the ROC curves for both unimodal systems and multimodal fusion using the Virtual database. At first glance, it can be observed that among all density-estimation-based techniques tested, the KDE-Prod and the MoG provided the best results in all experiments.

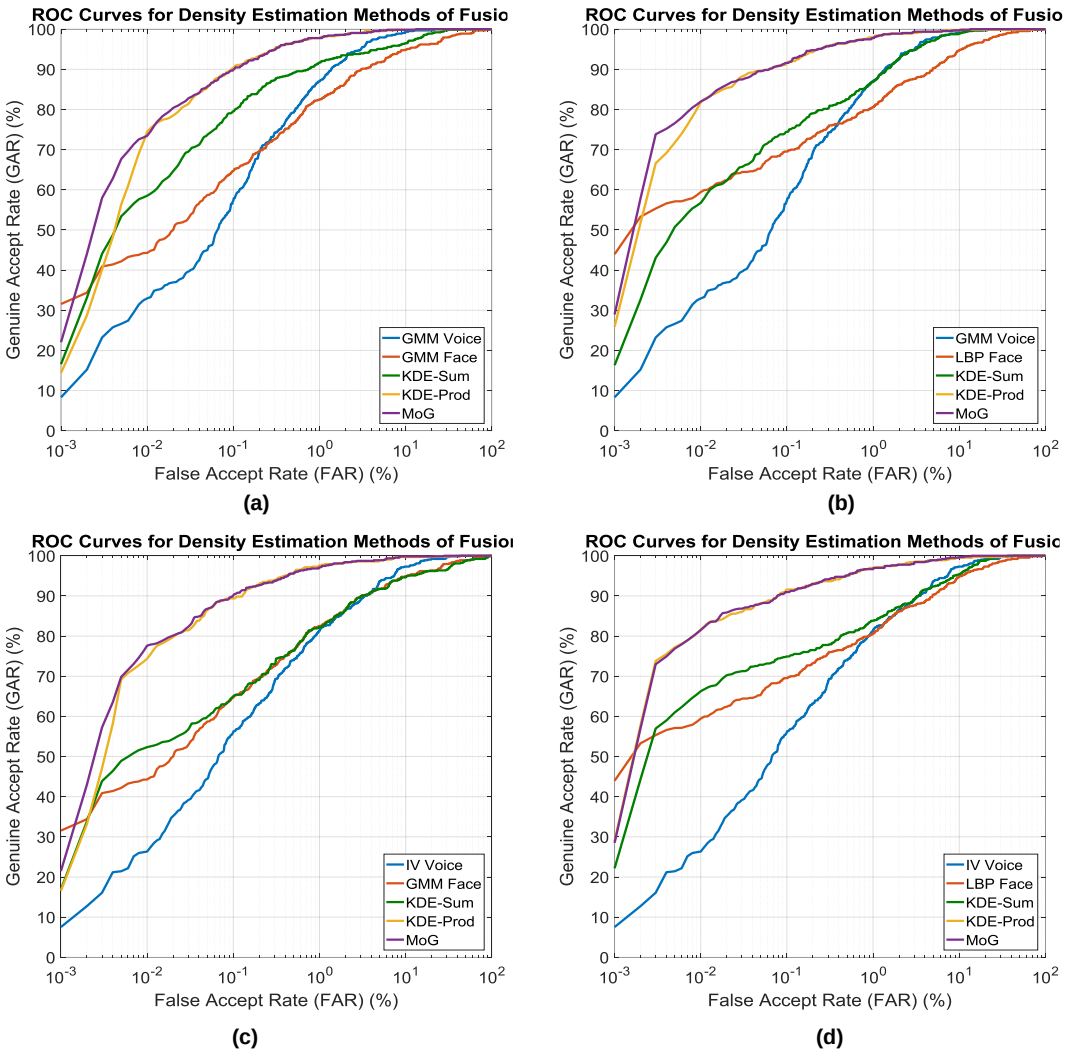

Figure 5-3: ROC Curves for Density-based Score Fusion in Virtual database. (a) Fusion of GMM/UBM Voice and GMM/UBM Face. (b) Fusion of GMM/UBM Voice and LBP Face. (c) Fusion of I-Vector Voice and GMM/UBM Face. (d) Fusion of I-Vector Voice and LBP Face.

More specifically, the Product of marginal density estimates using KDE delivered the lowest EER in three out of four combinations, although when compared with the MoG results, they are very similar, as it can be observed from the ROC Curves and Table 5-4, which summarizes the results for this fusion technique using the Virtual database.

	Fusion Technique								
	KDE Sum			KDE Product			MoG		
	GAR	AUC	EER	GAR	AUC	EER	GAR	AUC	EER
GMM Voice - GMM Face	79,52	0,991	4,919	90,32	0,999	**1,317**	89,86	0,999	1,371
GMM Voice - LBP Face	74,44	0,995	3,858	91,61	0,998	1,526	91,59	0,999	**1,392**
IV Voice - GMM Face	65,01	0,974	6,647	89,37	0,998	**1,789**	90,11	0,998	1,817
IV Voice - LBP Face	74,92	0,987	6,595	91,59	0,997	**2,247**	90,95	0,998	2,363

Table 5-4: Genuine Acceptance Rate (GAR), Area under ROC Curve (AUR) and Equal Error Rate (EER) of the density-based fusion methods for all classifiers combinations using the Virtual database.

5.2.2.2. MOBIO Database Results

In the MOBIO evaluation using the density-based methods, the results of the fusion are also better than the unimodal systems individually, as can be observed in Figure 5-4.

	Fusion Technique								
	KDE Sum			KDE Product			MoG		
	GAR	AUC	EER	GAR	AUC	EER	GAR	AUC	EER
GMM Voice - GMM Face	43,32	0,955	12,456	52,64	0,983	**6,737**	52,83	0,983	6,779
GMM Voice - LBP Face	34,70	0,956	11,412	50,56	0,985	**6,230**	50,05	0,985	6,288
IV Voice - GMM Face	34,01	0,911	16,774	51,23	0,988	5,622	51,87	0,988	**5,590**
IV Voice - LBP Face	35,60	0,943	14,899	52,21	0,990	4,939	52,42	0,990	**4,929**

Table 5-5: Genuine Acceptance Rate (GAR), Area under ROC Curve (AUR) and Equal Error Rate (EER) of the density-based fusion methods for all classifiers combinations using the MOBIO database.

Here, the KDE-Prod and MOG showed again the best result, with almost the same recognition performance between them.

In Table 5-5 are summarized the results for this fusion method using the MOBIO database. Here once again, it can be observed the similarity between the results obtained from the Product of KDE estimates and the MoG methods.

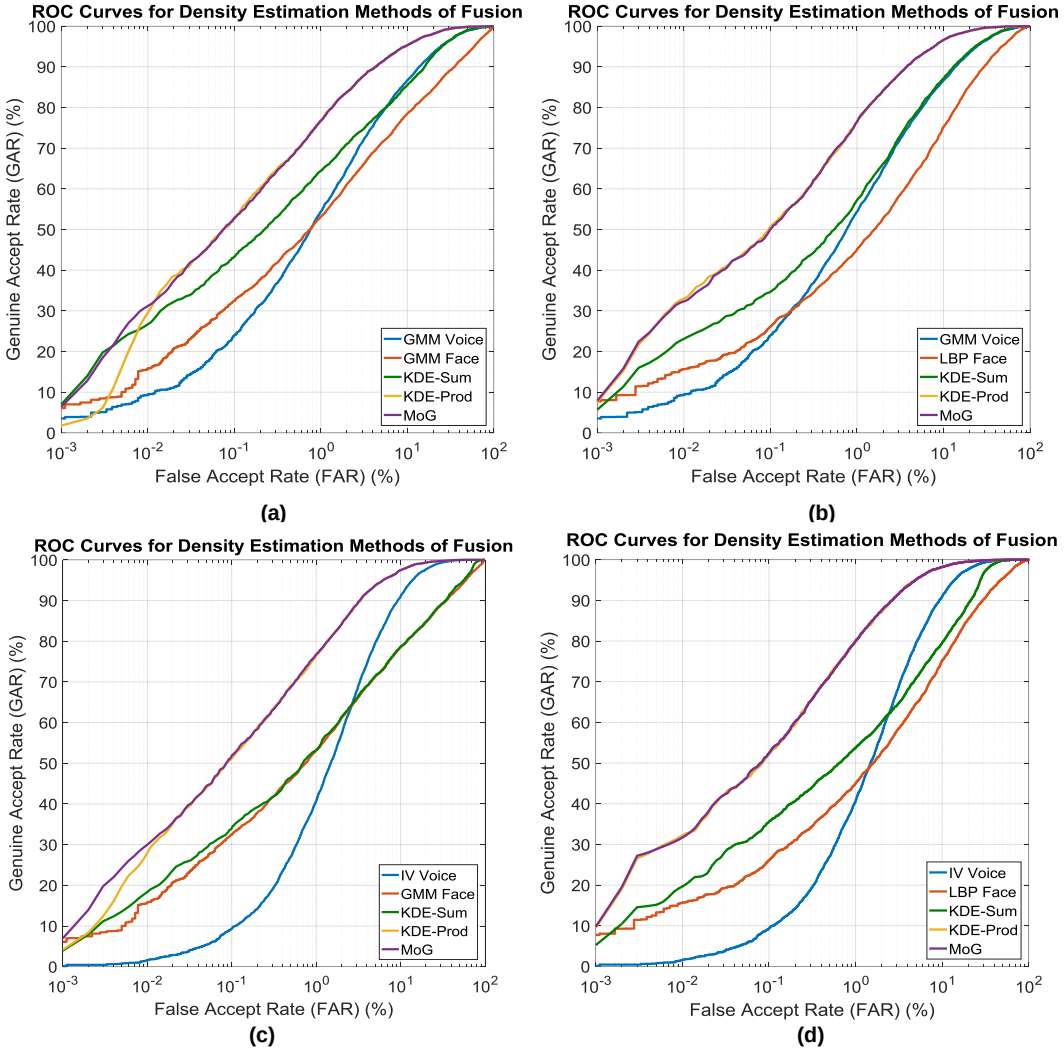

Figure 5-4: ROC Curves for Density-based Score Fusion in MOBIO database. (a) Fusion of GMM/UBM Voice and GMM/UBM Face. (b) Fusion of GMM/UBM Voice and LBP Face. (c) Fusion of I-Vector Voice and GMM/UBM Face. (d) Fusion of I-Vector Voice and LBP Face.

5.2.2.3. Classifier-based Fusion

Finally, the Classifier-Based Score Fusion approach was tested, using the output scores from the individual biometric systems as new features to train a probabilistic Support Vector Machine (SVM) and a probabilistic Random Forest

(RF) classifiers. The implementation of the SVM algorithm was based on the LibSVM public library [92].

One downside of classifier-based methods is that it is difficult to compute GAR values at specific desired FAR values, and thus it becomes hard to construct the ROC curves, because the classifiers are typically based on fixed thresholds in order to make the binary decisions. In our case for instance, for probabilistic versions, the class that has a posterior probability higher than 0.5 results in the predicted class. For that reason, in this work, we only computed the average GAR and average FAR on the five executions, instead of plotting the entire ROC curve, as in the previous experimental methods.

The parameters for the SVM classifier were selected following the configuration in [92], using a RBF kernel. The C and γ values are estimated using 3-fold cross-validation over the training set. In the case of the RF classifier, an ensemble of 500 trees was used in the forest, and the square root of the number of variables was used as the random variable subset.

To train the classifiers, some impostor scores were randomly discarded in order to balance the training process with the same amount of genuines and impostors.

5.2.2.4. Virtual Database Results

The results of this approach using the Virtual database are summarized in Table 5-6. Both methods of classification achieved very similar results, each of them obtaining the best EERs in two out of four configurations. Besides, the recognition accuracies for both are good in all combinations. Comparing the GAR values of this approach of fusion with GARs from the previous approaches in the ROC figures at the corresponding FAR values, it can be noted that the classification-based methods offers excellent performance. For instance, the GMM Voice and GMM Face configuration using MoG method achieved an average GAR of 98,57% at a FAR of 1.66%, while the SVM obtained 98.41%. In addition, at a FAR of 2%, the MoG provided 98.73% whereas the RF obtained 98.57%.

	Fusion Technique							
	SVM				RF			
	OA	EER	Avg FAR	Avg GAR	OA	EER	Avg FAR	Avg GAR
GMM Voice - GMM Face	98,34	**1,435**	1,66	98,41	98,00	1,552	2,00	98,57
GMM Voice - LBP Face	98,61	**1,714**	1,39	97,94	98,03	2,138	1,97	98,10
IV Voice - GMM Face	96,04	4,152	3,95	94,13	98,06	**2,056**	1,94	97,46
IV Voice - LBP Face	97,82	2,613	2,17	96,51	97,65	**2,055**	2,35	98,25

Table 5-6: Classification Overall Accuracy (OA), Equal Error Rate (EER), Average FAR (Avg FAR) and Average GAR (Avg GAR) of SVM and RF methods for each configuration using Virtual database.

5.2.2.5. MOBIO Database Results

For this database, the results of the classifier-based scores fusion method are presented in Table 5-7. Here the SVM classifier obtained the minimum EER for every configuration, being the last configuration the one with the lowest value. Compared with the other database, the Overall Accuracy is not as good, but in general the classifiers obtained good classification accuracies, above 90% in all cases.

	Fusion Technique							
	SVM				RF			
	OA	EER	Avg FAR	Avg GAR	OA	EER	Avg FAR	Avg GAR
GMM Voice - GMM Face	93,37	**6,854**	6,62	92,77	92,51	8,044	7,46	91,47
GMM Voice - LBP Face	93,37	**6,385**	6,64	93,76	92,54	7,237	7,47	92,97
IV Voice - GMM Face	93,16	**6,128**	6,88	94,88	92,94	6,260	7,09	94,26
IV Voice - LBP Face	94,20	**5,049**	5,84	95,83	93,91	5,634	6,12	95,06

Table 5-7: Classification Overall Accuracy (OA), Equal Error Rate (EER), Average FAR (Avg FAR) and Average GAR (Avg GAR) of SVM and RF methods for each configuration using MOBIO database.

5.3. Discussion

After evaluating the fusion approaches proposed in this work with all possible combination of matchers, some conclusions can be extracted.

In general, every proposed multimodal fusion technique yielded better results than the best unimodal biometric system, in terms of lower Equal Error Rates and higher Genuine Acceptance Rates and Classification Accuracy. In particular, between all fusion approaches tested, the density-based approaches attained the highest values of GAR and the minimum values of EER. In Figure 5-5 are shown the configurations of individual classifiers that provided the minimum value of EER for both databases, which is the Product of likelihood ratio using KDE for the Virtual database and the Mixture of Gaussians for the MOBIO.

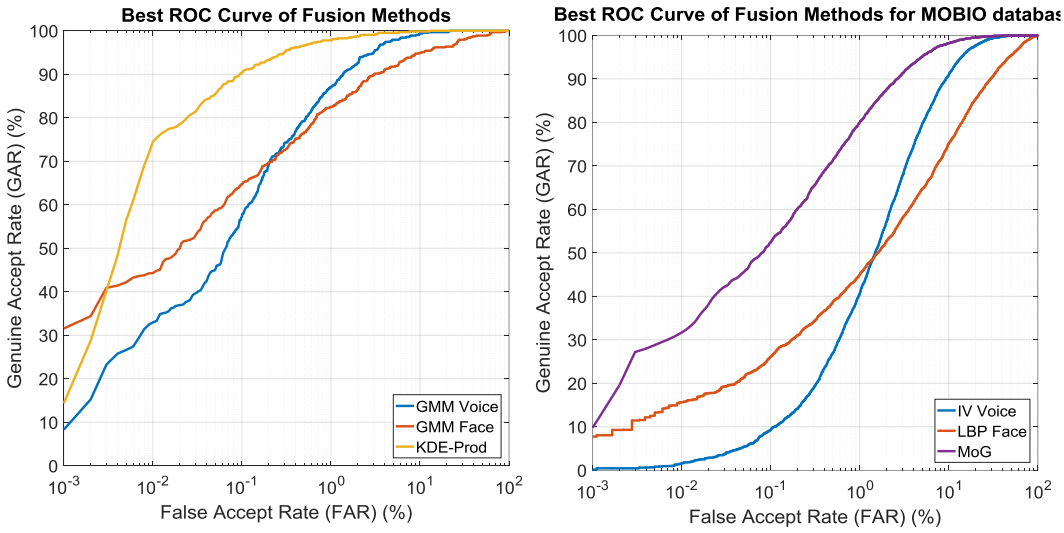

Figure 5-5: Best Configurations for Virtual database and MOBIO database

The results obtained in this work confirm other studies carried out to find the best fusion scheme among several tested, using different biometric modalities [44]. These studies concluded that density estimation-based approaches are the most accurate methods of fusion, but they are complex to implement as they rely on the estimation of the probability density functions, which is intrinsically a complex process.

It is important to remark that, in this work, the MoG method was preferable over the KDE because it was easier to implement and did not require any parameter tuning process.

Simple methods like Minmax normalization and Average Sum on the other hand provided very good results in our experiments in the transformation-based approaches. These methods are very convenient to use, given their simplicity to implement and their computational low cost.

Finally, the classifier approaches also shown good classification accuracy, being the SVM the best between the two.

6
CONCLUSIONS AND FUTURE WORKS

In this work, we have compared several techniques for score-level fusion in a bimodal system that combine independent speaker and facial recognition systems. For this purpose we tested four different combinations of matchers, two speaker recognition systems and two facial recognition systems.

The experimental analysis followed on Chapter 4 for configuring and assessing the performance of the unimodal systems, under verification and identification tasks, using a Virtual bimodal database and a real bimodal database, suggested that in applications with controlled conditions, these unimodal biometric matchers provide reasonable accuracy rates, sufficient for real scenarios. This was confirmed by the good results obtained using a Virtual database formed by two well-controlled databases of speeches and faces, such as TIMIT and FERET databases respectively. As the conditions deteriorate, and variations and noise start to appear in the acquired data, the recognition systems suffered to preserve acceptable recognition rates. With the use of a more challenging database such as MOBIO, we could observe that behavior.

In Chapter 5, we investigated three main approaches for combining the scores output by the unimodal classifiers in order to overcome their intrinsic limitations under more adverse conditions. These approaches are the Transformation-based score fusion, the Density-based score fusion and the Classifier-based score fusion.

By comparing throughout our experiments the ROC Curves, the Equal Error Rate (EER), the Area under ROC curve (AUC) and the Genuine Acceptance Rates (GAR) metrics for each of these approaches, we concluded that the Density-based score fusion provided the most consistent results among all configurations of bimodal fusion.

Among the Transformation-based approach, the Minmax normalization scheme with the Average Sum rule exhibited a very good performance, being behind the density-based methods by a small margin in every tested configuration. The Classifier-based methods on the other hand also performed reasonable good,

but these methods have the limitation that they cannot provide high GARs at low FARs, and cannot be compared directly with the other methods using ROC curves.

Finally, the good multimodal results obtained from the Virtual database indicate that the initial statistical independence assumption made between the voice and facial biometric traits was reasonable.

A desirable extension of this work would include the configuration of the parameters for the unimodal biometric systems using the MOBIO database, because the parameters in this work were fixed for the MOBIO database, in contrast with the Virtual database, that it was completely parameterized.

It would also be interesting to further investigate the multialgorithm and multimodal fusion between the implemented unimodal systems to see if this approach provide better recognition accuracy and higher values of GAR. Another possibility not covered in this work is the inclusion of the identification operating mode, since all experiments were tested only for verification mode.

REFERENCES

1. Nandakumar, K., *Multibiometric systems: Fusion strategies and template security*. 2008: ProQuest.

2. Bigun, E.S., et al., *Expert conciliation for multi modal person authentication systems by Bayesian statistics.* Audio- and Video-Based Biometric Person Authentication, 1997. **1206**: p. 291-300.

3. Verlinde, P., G. Chollet, and M. Acheroy, *Multi-modal identity verification using expert fusion.* Information Fusion, 2000. **1**(1): p. 17-33.

4. Kittler, J., et al., *On combining classifiers.* Ieee Transactions on Pattern Analysis and Machine Intelligence, 1998. **20**(3): p. 226-239.

5. Ben-Yacoub, S., Y. Abdeljaoued, and E. Mayoraz, *Fusion of face and speech data for person identity verification.* IEEE Trans Neural Netw, 1999. **10**(5): p. 1065-74.

6. Brunelli, R. and D. Falavigna, *Person Identification Using Multiple Cues.* Ieee Transactions on Pattern Analysis and Machine Intelligence, 1995 **17**(10): p. 955-966.

7. Ross, A. and A. Jain, *Information fusion in biometrics.* Pattern Recognition Letters, 2003. **24**(13): p. 2115-2125.

8. Hong, L. and A. Jain, *Integrating faces and fingerprints for personal identification.* Ieee Transactions on Pattern Analysis and Machine Intelligence, 1998. **20**(12): p. 1295-1307.

9. Ho, T.K., J.J. Hull, and S.N. Srihari, *Decision Combination in Multiple Classifier Systems.* Ieee Transactions on Pattern Analysis and Machine Intelligence, 1994. **16**(1): p. 66-75.

10. Prabhakar, S. and A.K. Jain, *Decision-level fusion in fingerprint verification.* Pattern Recognition, 2002. **35**(4): p. 861-874.

11. Verlinde, P. *Comparing decision fusion paradigms using k-NN based classifiers, decision trees and logistic regression in a multi-modal identity verification application*. 1999. Citeseer.

12. Ben-Yacoub, S., Y. Abdeljaoued, and E. Mayoraz, *Fusion of face and speech data for person identity verification.* Ieee Transactions on Neural Networks, 1999. **10**(5): p. 1065-1074.

13. Hanmandlu, M., et al., *Score level fusion of multimodal biometrics using triangular norms.* Pattern Recognition Letters, 2011. **32**(14): p. 1843-1850.

14. Messer, K., et al. *XM2VTSDB: The extended M2VTS database*. in *Second international conference on audio and video-based biometric person authentication*. 1999. Citeseer.

15. Bailly-Bailliere, E., et al., *The BANCA database and evaluation protocol.* Audio-and Video-Based Biometric Person Authentication, Proceedings, 2003. **2688**: p. 625-638.

16. Garcia-Salicetti, S., et al., *BIOMET: A multimodal person authentication database including face, voice, fingerprint, hand and signature modalities.* Audio-and Video-Based Biometric Person Authentication, Proceedings, 2003. **2688**: p. 845-853.

17. Fierrez, J., et al., *BiosecurID: a multimodal biometric database.* Pattern Analysis and Applications, 2010. **13**(2): p. 235-246.

18. McCool, C., et al., *Bi-Modal Person Recognition on a Mobile Phone: using mobile phone data.* 2012 Ieee International Conference on Multimedia and Expo Workshops (Icmew), 2012: p. 635-640.

19. Hicklin, A., B. Ulery, and C. Watson, *A brief introduction to biometric fusion.* National institute of standards and technology, 2006.

20. Liu, X.M. and T. Chen, *Geometry-assisted statistical modeling for face mosaicing.* 2003 International Conference on Image Processing, Vol 2, Proceedings, 2003: p. 883-886.

21. Yang, F., et al., *Development of a fast panoramic face mosaicking and recognition system.* Optical Engineering, 2005. **44**(8).

22. Choi, K., H. Choi, and J. Kim, *Fingerprint mosaicking by rolling and sliding.* Audio and Video Based Biometric Person Authentication, Proceedings, 2005. **3546**: p. 260-269.

23. Zhang, Y.L., J. Yang, and H.T. Wu, *A hybrid swipe fingerprint mosaicing scheme.* Audio and Video Based Biometric Person Authentication, Proceedings, 2005. **3546**: p. 131-140.

24. Jain, A. and A. Ross, *Fingerprint mosaicking.* 2002 Ieee International Conference on Acoustics, Speech, and Signal Processing, Vols I-Iv, Proceedings, 2002: p. 4064-4067.

25. Moon, Y.S., et al., *Template synthesis and image mosaicking for fingerprint registration: An experimental study.* 2004 Ieee International Conference on Acoustics, Speech, and Signal Processing, Vol V, Proceedings, 2004: p. 409-412.

26. Singh, S., et al., *Infrared and visible image fusion for face recognition.* Biometric Technology for Human Identification, 2004. **5404**: p. 585-596.

27. Wang, J., et al., *Performance evaluation of infrared and visible image fusion algorithms for face recognition.* Proceedings of the International Conference on Intelligent Systems and Knowledge Engineering (Iske 2007), 2007.

28. Heo, J., et al. *Fusion of visual and thermal signatures with eyeglass removal for robust face recognition*. in *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*. 2004. IEEE.

29. Ross, A. and R. Govindarajan, *Feature level fusion using hand and face biometrics.* Biometric Technology for Human Identification II, 2005. **5779**: p. 196-204.

30. Chibelushi, C.C., J.S. Mason, and F. Deravi, *Integrated person identification using voice and facial features.* IEEE CIRCUITS AND DEVICES, 1992. **1**(74): p. 4-4.

31. Chibelushi, C.C., J.S.D. Mason, and F. Deravi, *Feature-level data fusion for bimodal person recognition.* Sixth International Conference on Image Processing and Its Applications, Vol 1, 1997(443): p. 399-403.

32. Shah, D., K.J. Han, and S.S. Nayaranan, *A Low-Complexity Dynamic Face-Voice Feature Fusion Approach to Multimodal Person Recognition.* 2009 11th Ieee International Symposium on Multimedia (Ism 2009), 2009: p. 24-31.

33. Son, B. and Y. Lee, *Biometric authentication system using reduced joint feature vector of iris and face.* Audio and Video Based Biometric Person Authentication, Proceedings, 2005. **3546**: p. 513-522.

34. Gan, J.Y. and Y. Liang, *A method for face and iris feature fusion in identity authentication.* Int. J. Comp. Sci. Netw. Secur, 2006. **6**(2): p. 135-138.

35. Chen, C.-H. and C. Te Chu, *Fusion of face and iris features for multimodal biometrics*, in *Advances in Biometrics*. 2006, Springer. p. 571-580.

36. Wang, Z., et al., *Feature-level fusion of iris and face for personal identification*, in *Advances in Neural Networks–ISNN 2009*. 2009, Springer. p. 356-364.

37. Kumar, A., et al., *Personal verification using palmprint and hand geometry biometric.* Audio-Based and Video-Based Biometric Person Authentication, Proceedings, 2003. **2688**: p. 668-678.

38. Gao, Y. and M. Maggs, *Feature-level fusion in personal identification.* 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol 1, Proceedings, 2005: p. 468-473.

39. Zhon, X.L. and B. Bhanu, *Integrating face and gait for human recognition at a distance in video.* Ieee Transactions on Systems Man and Cybernetics Part B-Cybernetics, 2007. **37**(5): p. 1119-1137.

40. Jain, A.K. and B. Chandrasekaran, *39 Dimensionality and sample size considerations in pattern recognition practice.* Handbook of statistics, 1982. **2**: p. 835-855.

41. Duc, B., et al., *Person authentication by fusing face and speech information.* Audio- and Video-Based Biometric Person Authentication, 1997. **1206**: p. 311-318.

42. Pigeon, S. and L. Vandendorpe, *The M2VTS multimodal face database (Release 1.00).* Audio- and Video-Based Biometric Person Authentication, 1997. **1206**: p. 403-409.

43. Verlinde, P., et al., *Applying Bayes based classifiers for decision fusion in a multi-modal identity verification system.* 1999.

44. Ulery, B., et al., *Studies of biometric fusion*. 2006: US Department of Commerce, National Institute of Standards and Technology.

45. Nandakumar, K., et al., *Likelihood ratio-based biometric score fusion.* Ieee Transactions on Pattern Analysis and Machine Intelligence, 2008. **30**(2): p. 342-347.

46. Verlinde, P. and G. Chollet. *Combining vocal and visual cues in an identity verification system using k-nn based classifiers*. in *Multimedia Signal Processing, 1998 IEEE Second Workshop on*. 1998. IEEE.

47. Fierrez-Aguilar, J., et al., *A comparative evaluation of fusion strategies for multimodal biometric verification.* Audio-and Video-

Based Biometric Person Authentication, Proceedings, 2003. **2688**: p. 830-837.

48. Ortega-Garcia, J., et al., *MCYT baseline corpus: a bimodal biometric database.* Iee Proceedings-Vision Image and Signal Processing, 2003. **150**(6): p. 395-401.

49. Wang, F. and J. Han, *Multimodal biometric authentication based on score level fusion using support vector machine.* Opto-Electronics Review, 2009. **17**(1): p. 59-64.

50. *The ORL Database of Faces*. [Web Page] 2002 AT&T Laboratories Cambridge]. Available from: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.

51. Proenca, H. and L.A. Alexandre, *UBIRIS: A noisy iris image database.* Image Analysis and Processing - Iciap 2005, Proceedings, 2005. **3617**: p. 970-977.

52. Ma, Y., B. Cukic, and H. Singh, *A classification approach to multi-biometric score fusion.* Audio and Video Based Biometric Person Authentication, Proceedings, 2005. **3546**: p. 484-493.

53. Breiman, L., *Random forests.* Machine Learning, 2001. **45**(1): p. 5-32.

54. Wang, Y.H., T.N. Tan, and A.K. Jain, *Combining face and iris biometrics for identity verification.* Audio-and Video-Based Biometric Person Authentication, Proceedings, 2003. **2688**: p. 805-813.

55. Jain, A., K. Nandakumar, and A. Ross, *Score normalization in multimodal biometric systems.* Pattern Recognition, 2005. **38**(12): p. 2270-2285.

56. Snelick, R., et al., *Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems.* Ieee Transactions on Pattern Analysis and Machine Intelligence, 2005. **27**(3): p. 450-455.

57. Tao, Q., *Face verification for mobile personal devices.* 2009.

58. Daugman, J. *Combining Multiple Biometrics* [Web Page]; The Computer Laboratory, Cambridge University]. Available from: http://www.cl.cam.ac.uk/users/jgd1000/combine/combine.html.

59. Lam, L. and C.Y. Suen, *Application of majority voting to pattern recognition: An analysis of its behavior and performance.* Ieee Transactions on Systems Man and Cybernetics Part a-Systems and Humans, 1997. **27**(5): p. 553-568.

60. Kuncheva, L.I., *Combining pattern classifiers: methods and algorithms*. 2004: John Wiley & Sons.

61. Xu, L., A. Krzyzak, and C. Suen, *ªMethods for Combining Multiple Classifiers and Their Applications in Handwritten Character Recognition.* º IEEE Trans. Systems, Man, and Cybernetics, 1992. **22**: p. 418-435.

62. Huang, Y.S. and C.Y. Suen, *Method of Combining Multiple Experts for the Recognition of Unconstrained Handwritten Numerals.* Ieee Transactions on Pattern Analysis and Machine Intelligence, 1995. **17**(1): p. 90-94.

63. Togneri, R. and D. Pullella, *An Overview of Speaker Identification: Accuracy and Robustness Issues.* Ieee Circuits and Systems Magazine, 2011. **11**(2): p. 23-61.

64. Kinnunen, T., *Spectral Features for Automatic Text-Independent Speaker Recognition*, in *Department of Computer Science*. 2004 University of Joensuu: Finland.

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

66. Kinnunen, T. and H. Li, *An overview of text-independent speaker recognition: From features to supervectors.* Speech Communication, 2010. **52**(1): p. 12-40.

67. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the royal statistical society. Series B (methodological), 1977: p. 1-38.

68. Kenny, P., G. Boulianne, and P. Dumouchel, *Eigenvoice modeling with sparse training data.* Ieee Transactions on Speech and Audio Processing, 2005. **13**(3): p. 345-354.

69. Dehak, N., et al., *Front-End Factor Analysis for Speaker Verification.* Ieee Transactions on Audio Speech and Language Processing, 2011. **19**(4): p. 788-798.

70. Matrouf, D., et al., *A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification.* Interspeech 2007: 8th Annual Conference of the International Speech Communication Association, Vols 1-4, 2007: p. 2624-2627.

71. Prince, *Probabilistic linear dsicriminant analisys for inferences about identity.pdf.* 2007

72. Kenny, P. *Bayesian Speaker Verification with Heavy-Tailed Priors*. in *Odyssey*. 2010.

73. Garcia-Romero, D. and C.Y. Espy-Wilson, *Analysis of I-vector Length Normalization in Speaker Recognition Systems.* 12th Annual Conference of the International Speech Communication Association 2011 (Interspeech 2011), Vols 1-5, 2011: p. 256-259.

74. Sanderson, C. and K.K. Taliwal, *Polynomial features for robust face authentication.* 2002 International Conference on Image Processing, Vol Iii, Proceedings, 2002: p. 997-1000.

75. Ojala, T., M. Pietikäinen, and D. Harwood, *A comparative study of texture measures with classification based on feature distributions.* Pattern Recognition, 1996. **29**(1): p. 51-59.

76. Feitosa, R.Q., et al., *Weighting Estimation for Texture-Based Face Recognition Using the Fisher Discriminant.* Computing in Science & Engineering, 2011. **13**(3): p. 31-37.

77. Ahonen, T., A. Hadid, and M. Pietikainen, *Face recognition with local binary patterns.* Computer Vision - Eccv 2004, Pt 1, 2004. **3021**: p. 469-481.

78. Lehmann, E.L., J.P. Romano, and G. Casella, *Testing statistical hypotheses*. Vol. 150. 1986: Wiley New York et al.

79. Figueiredo, M.A.T. and A.K. Jain, *Unsupervised learning of finite mixture models.* Ieee Transactions on Pattern Analysis and Machine Intelligence, 2002. **24**(3): p. 381-396.

80. Cortes, C. and V. Vapnik, *Support-Vector Networks.* Machine Learning, 1995. **20**(3): p. 273-297.

81. Hastie, T., R. Tibshirani, and J.H. Friedman, *The elements of statistical learning : data mining, inference, and prediction*. 2nd ed. Springer series in statistics. 2009, New York: Springer. xxii, 745 p.

82. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. [Web Page] 1993; Linguistic data consortium]. Available from: https://catalog.ldc.upenn.edu/LDC93S1.

83. *Face Recognition Technology (FERET)*. [Web Page] 1996; National Institute of Standards and Technology (NIST)]. Available from: http://www.nist.gov/itl/iad/ig/feret.cfm.

84. Steeneken, H.J.M. and A. Varga, *Assessment for Automatic Speech Recognition .1. Comparison of Assessment Methods.* Speech Communication, 1993. **12**(3): p. 241-246.

85. Varga, A. and H.J.M. Steeneken, *Assessment for Automatic Speech Recognition .2. Noisex-92 - a Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems.* Speech Communication, 1993. **12**(3): p. 247-251.

86. *Samples of noisex-92 database*. [Web Page] 2013; Signal Processing Information Base (SPIB)]. Available from: http://spib.linse.ufsc.br/noise.html.

87. Moon, H. and P.J. Phillips, *Computational and performance aspects of PCA-based face-recognition algorithms.* Perception, 2001. **30**(3): p. 303-321.

88. Khoury, E., et al., *The 2013 Speaker Recognition Evaluation in Mobile Environment.* 2013 International Conference on Biometrics (Icb), 2013.

89. Gunther, M., et al., *The 2013 Face Recognition Evaluation in Mobile Environment.* 2013 International Conference on Biometrics (Icb), 2013.

90. McCool, C., et al., *Session variability modelling for face authentication.* Iet Biometrics, 2013. **2**(3): p. 117-129.

91. Lucey, S. and T. Chen. *A GMM parts based face representation for improved verification through relevance adaptation*. in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*. 2004. IEEE.

92. Chang, C.C. and C.J. Lin, *LIBSVM: A Library for Support Vector Machines.* Acm Transactions on Intelligent Systems and Technology, 2011. **2**(3).