UNIVERSITÀ DEGLI STUDI DI CATANIA

Dipartimento di Ingegneria Elettrica, Elettronica ed Informatica

Dottorato di Ricerca in Ingegneria Informatica e delle Telecomunicazioni

XXIV Ciclo

INNOVATIVE ALGORITHMS, TRAITS AND APPLICATION SCENARIOS FOR MONO-MULTIMODAL BIOMETRIC RECOGNITION

Ing. Andrea Spadaccini

Coordinatore	Tutor
Chiar.mo Prof. O. Mirabella	Chiar.mo Prof. F. Beritelli

to Rita, Nico, Giorgia – my family

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Identity verification is one of the most common and important processes in our daily lives. For centuries, humans have relied on the visual appearance of their peers - or other distinctive traits - to recognize them.

Currently, most of the traditional authentication methods infer one's identity by either verifying the knowledge of a shared secret (i.e., password) or the possession of a given object (i.e., token); both methods are sub-optimal, since they do not directly verify the person's identity, but an alternate and less rich representation that could also very easily stolen or inadvertently shared.

Biometrics offers a natural solution to this problem, by providing quantitative methods for recognizing one's identity by the analysis of either physiological or behavioural traits.

In addition to the traditional biometric traits, such as fingerprint, iris or voice, there is a growing interest in novel biometric traits, that can be used in conjunction with the most established ones to compensate to their weaknesses. In the first part of this thesis, we will analyze the usage of heart sounds as a physiological trait for biometric recognition, discussing some novel ad-hoc algorithms developed to process them.

Forensic analysts often have to determine whether a given speech sam-

ple was uttered by a suspect or not; in the second part of the thesis, we will investigate the usage of automatic text-independent speaker recognition systems in this context, exploring the limits of this approach and proposing new solutions.

Finally, given the capillary diffusion that Internet access has gained in the last years, we will analyze the problem of biometric authentication for web application. Through the performance analysis of 3 biometric systems and their combination using 2 multi-biometric score level fusion strategies, we will find the optimal combination of those system; we will then present the architecture and implementation of an open-source web-based multi-biometric authentication system based on speech and face recognition, fused together using the optimal strategy identified in the preliminary analysis phase.

SOMMARIO

Provare la propria identità è una delle attività più comuni ed importanti che ci viene richiesto di svolgere quotidianamente. Nei secoli, gli esseri umani si sono affidati alla fisionomia dei propri simili - o ad altri tratti distintivi - per riconoscerli. Attualmente, la maggior parte dei metodi di autenticazione derivano l'identità di una persona tramite la verifica di un segreto condiviso (come una password) o tramite la verifica del possesso di un determinato oggetto (un *token*); entrambi i metodi non sono esenti da difetti, poiché non verificano direttamente l'identità della persona, ma una rappresentazione della stessa alternativa e meno significativa, che potrebbe inoltre essere rubata o inavvertitamente condivisa.

I metodi di autenticazione biometrica offrono una valida soluzione a questo problema, fornendo metodi quantitativi per il riconoscimento dell'identità di un individuo tramite l'analisi di tratti fisiologici o comportamentali.

Uno degli obiettivi principali delle attuali attività di ricerca nell'ambito della biometria è l'individuazione di nuovi tratti biometrici, che possano essere utilizzati in aggiunta ai tratti biometrici tradizionali per compensarne le debolezze. Nella prima parte di questa tesi verrà analizzato l'utilizzo dei suoni

cardiaci come tratti fisiologici per il riconoscimento biometrico, illustrando algoritmi e sistemi innovativi che sfruttano questa tecnica.

Nella seconda parte della tesi si analizzerà l'utilizzo di tecniche automatiche di riconoscimento biometrico basate sulla voce in contesto forense. La qualità del segnale audio analizzato in questi contesti è spesso non eccellente a causa del rumore introdotto dal processo di acquisizione e dall'ambiente in cui viene effettuata la registrazione, che non sono sotto il controllo del perito. In queste difficili condizioni, attualmente il perito viene chiamato ad effettuare delle scelte che potrebbero condizionare l'esito del riconoscimento e di conseguenza anche influenzare l'esito del processo stesso. L'analisi presentata si focalizza sulle limitazioni della biometria vocale in questo contesto, e presenta delle nuove soluzioni per aggirare le stesse.

Infine, considerato il continuo incremento della diffusione dell'accesso ad Internet e la crescente importanza che le attività basate sul web stanno assumendo negli ultimi anni, nella terza parte della tesi sarà analizzato il tema dell'autenticazione biometrica nell'ambito delle applicazioni web. Tramite l'analisi delle prestazioni di 3 sistemi biometrici e le loro combinazioni utilizzando 2 strategie di fusione multi-biometrica a livello di punteggio, sarà identificata la combinazione ottimale degli stessi; saranno quindi presentate e discusse sia l'architettura che l'implementazione di un sistema open-source di autenticazione multi-biometrica per applicazioni web basate su riconoscimento del volto e della voce, fusi tramite la strategia ottimale identificata nella fase di analisi preliminare.

CHAPTER ONE

INTRODUCTION

Identity verification is an increasingly important process in our daily lives. Whether we need to use our own equipment or to prove our identity to third parties in order to use services or gain access to physical places, we are constantly required to declare our identity and prove our claim.

Traditional authentication methods fall into two categories: proving that you know something (i.e., password-based authentication) and proving that you own something (i.e., token-based authentication).

These methods connect the identity with an alternate and less rich representation, for instance a password, that can be lost, stolen, or shared.

A solution to these problems comes from biometric recognition systems. Biometrics offers a natural solution to the authentication problem, as it contributes to the construction of systems that can recognize people by the analysis of their physiological and/or behavioral characteristics. With biometric systems, the representation of the identity is something that is directly derived from the subject, therefore it has properties that a surrogate representation, like a password or a token, simply cannot have [1–3]. Biometric recognition is further discussed in Chapter 2.

1.1 Novel traits, algorithms and application scenarios

In this thesis, we will describe our work in some of the most challenging research areas in the field of biometric recognition: novel biometric traits, novel algorithms for biometric recognition, novel application scenarios for biometrics.

One of the most important research directions in the field of biometrics is the characterization of novel biometric traits that can be used in conjunction with other traits, to limit their shortcomings or to enhance their performance. In Chapter 3 we will describe our proposal for a new physiological biometric trait: the heart sound.

Throughout the thesis, and especially in Chapter 3, we will present some novel algorithms invented during the research activities; amongst them, the most important are the ones for the computation of the First-to-Second Ratio (Section 3.3.2) and the quality-based best subsequence selection (Section 3.5.1), both designed for heart-sounds based biometry.

In this thesis we will also examine innovative and challenging application scenarios for biometric recognition systems. In Chapter 4 we deal with speaker recognition in the delicate context of forensic investigation. As of today, there is no definitive answer to the question of whether automatic speaker recognition can be successfully used in the forensic context, where speakers have to be identified from fragments of conversations captured in noisy conditions, like from wire-tappings or from ambient microphones. The analysis that we will present has the objective to find the current limitations of this technology with respect to most of the stages of speaker recognition systems that deal directly with the audio signals, and to try to make it clearer how these systems can be improved in order to be, in a near future, successfully employed for forensic tasks.

Finally, in Chapter 5 we will present a novel architecture for multi-modal biometric recognition in the context of web applications. The ensemble of classifiers and biometric traits used for this system have been determined by an experimental selection process involving testing all the possible combinations of 1-modal, 2-modal and 3-modal systems built using exploiting three different traits and two different score-level multi-biometric fusion strategies.

1.2 Research projects

Most of the research activity presented in this thesis was carried on in the framework of the following research projects:

- **ICT-E1**- Dipartimento di Ingegneria Informatica e delle Telecomunicazioni - University of Catania (2008-2009)
- **Interactive Multimedia Services** - Consorzio Nazionale Interuniversitario per le Telecomunicazioni (2009-2010)
- **Biometric4Net** - Consortium GARR (2010-2011)
- **Context-Aware Security by Hierarchical Multilevel Architectures (CASHMA)** - Centro di Competenza ICT-SUD and Engineering S.p.A. (2011)

CHAPTER

TWO

BIOMETRIC RECOGNITION

As stated in Chapter 1, identity verification is crucial to many of our daily activities. Let *P* be a person and *I* be an identity; there are two types of identification:

- Positive identification: verification that the stated identity claim is true (the identity of *P* is *I*);
- Negative identification: verification that a negative identity claim is true (the identity of *P* is **not** *I*).

While the first type is the most common, there are some use-cases also for the second kind, like preventing issuing multiple identity documents of the same type to the same person (think of *P* having documents for identities I_1 and I_2). For the rest of this thesis, unless otherwise specified, we will talk about identification in the sense of positive identification.

Traditionally, there are two ways of doing identity verification:

- verify that the person has something (a token, an ID card, etc..);

- verify that the person knows something (a password, a PIN, etc..).

Both these approaches reduce the complex problem of answering the question “Who are you?” (or “Is I your identity?”) to simpler problems: the first approach is equivalent to asking the question “Do you have the object X?”; the second is equivalent to the question “Do you know X?”.

Note that both the approaches can be decoupled from the person P. In other words, the person who is claiming identity I (that corresponds to person P) may not be P and yet be able to make a successful claim on identity I.

Biometric recognition aims to provide an answer to the question “Who are you?” that is intrinsically dependent on the properties of the person itself and not on external objects or pieces of knowledge; it answers this question by processing signals that derive either from characteristics of the human body (physiological traits) or from a given behaviour like walking (behavioural traits) [4, 5].

In this chapter, we will give a brief overview of biometric systems in Section 2.1; in Section 2.2 we will discuss about the performance metrics used to evaluate such systems; in Section 2.3 we will present a short description of the most important biometric traits; finally, in Section 2.4 we will introduce the concept of multibiometrics.

2.1 Biometric systems

A biometric system is basically a pattern recognition system, that acquires data from the user P, extracts features from the data, compares the feature set with one or more stored identity models (depending on the operating modality) and then performs the identity recognition task [6]. Figure 2.1 shows a diagram of a generic biometric system.

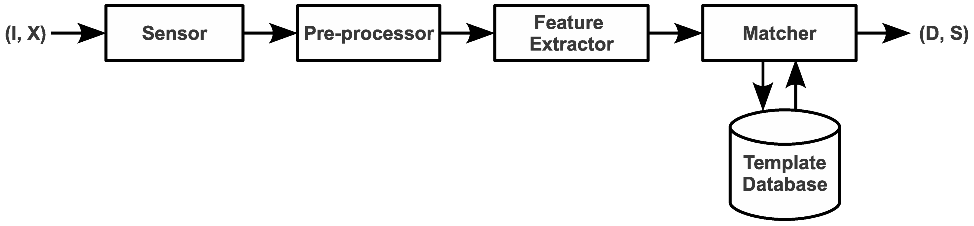


Figure 2.1: Diagram of a biometric system

Its five components are the following

1. **Sensor**, the device that is responsible for the acquisition of biometric data from the user;

2. **Pre-processor**, the part of the system responsible for preparing the raw data for the further processing steps; it might include, for instance, a filter in case of audio signals, or image enhancement algorithms when dealing with image-based biometric traits;

3. **Feature extractor**, the part of the system that is responsible for analyzing the acquired data and extracting from it an alternate representation, usually more meaningful and more compact; an example is the extraction of Mel-Frequency Cepstrum Coefficients (MFCC) in voice-based biometric systems;

4. **Template database**, the storage area that contains the models of all the enrolled identities;

5. **Matcher**, the module that has the duty to compare the feature set(s) extracted to the models, according to the operating modality of the bio-

metric system; the output of this stage is either a matching score or, in the case of simpler systems, a decision.

A biometric system can work in two modalities: enrollment and recognition.

Enrollment is the process by which an user, whose identity is verified by external means such as an ID card, deposits its biometric data in the system, so that it can build a template from this data and store it in the template database, associating it with the verified identity. The enrollment process does not use the Matcher component of the biometric system, and it is necessary if the same person needs to use the system in any other operating modalities.

After a user is enrolled in the system, the data related to his identity can be used during the recognition phase; there are two types of recognition: identification and verification.

Identification is the process by which only the biometric data is fed into the biometric system, and using only this data it has to determine which is (or which are) the most likely identity (resp. identities) in the database; this means doing a 1:N comparison.

Verification is the process by which both the biometric data and a claimed identity are fed to the system, that must check if the data matches with the template associated to the identity, doing a 1:1 comparison.

The system shown in Figure 2.1 is an example system that takes in input both the identity I and biometric data X – thus working in the verification modality – and outputs both the matching score S and the decision D.

In the rest of this thesis, unless otherwise specified, we will discuss about identification systems.

2.2 Performance of biometric systems

A biometric identity verification system can be seen as a binary classifier.

Binary classification systems work by comparing matching scores to a threshold; their accuracy is therefore closely linked with the choice of the threshold, which must be selected according to the context of the system.

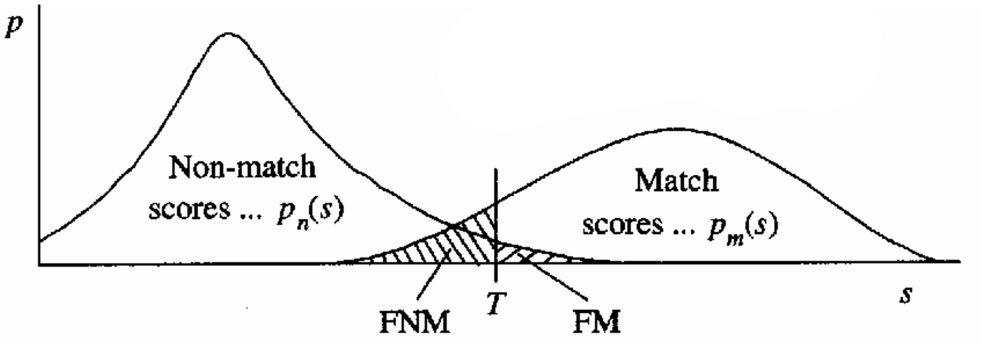


Figure 2.2: Biometric system errors

There are two possible errors that a binary classifier can make:

- **False Match (Type I Error):** accept an identity claim even if the template does not match with the model;

- **False Non-Match (Type II Error):** reject an identity claim even if the template matches with the model

The importance of each type of errors depends on the context in which the biometric system operates; for instance, in a high-security environment, a Type I error can be critical, while Type II errors could be tolerated.

Given a threshold T, and given a distribution of scores such as the one depicted in Figure 2.2, we can represent the two error types using the following formulas:

$$FNM(T) = \int_{-\infty}^{T} p_n(s)ds \tag{2.1}$$

$$FM(T) = \int_{T}^{+\infty} p_m(s)ds \tag{2.2}$$

where $p_n(s)$ is the distribution of non-match (impostor) scores and $p_m(s)$ is the distribution of match (genuine) scores, under the assumption that higher scores lead to a higher matching likelihood.

The choice of a threshold is a delicate design challenge that must be undertaken while working on a biometric system. Biometric systems are deployed in a wide range of working contexts, from consumer devices to military-grade access control, and each of them needs a careful trade-off between usability and security.

When evaluating the performance of a biometric system, however, we need to take a threshold-independent approach, because we cannot know its applications in advance. A common performance measure is the Equal Error Rate (EER) [4], defined as the error rate at which the False Match Rate (FMR) is equal to the False Non-Match Rate (FNMR).

A finer evaluation of biometric systems can be done by plotting the Detection Error Trade-off (DET) curve, that is the plot of FMR against FNMR. This allows to study their performance when a low FNMR or FMR is imposed to the system.

The DET curve represents the trade-off between security and usability. A system with low FMR is a highly secure one but will lead to more non-matches, and can require the user to try the authentication process several times; a system with low FNMR will be more tolerant and permissive, but will make more false match errors, thus letting more unauthorized users to get a positive match. The choice between the two setups, and between all the intermediate security levels, is strictly application-dependent.

2.3 Biometric traits

The authors of [1] present a classification of the most common biometric traits with respect to the 7 qualities that, according to them, are the most significant parameters to use for a meaningful and complete comparison. Those qualities of biometric traits are:

- **Universality**: each person should possess it;
- **Distinctiveness**: it should be helpful in the distinction between any two people;
- **Permanence**: it should not change over time;
- **Collectability**: it should be quantitatively measurable;
- **Performance**: biometric systems that use it should be reasonably performing, with respect to speed, accuracy and computational requirements;
- **Acceptability**: the users should see its usage as a natural and trustable thing to do in order to authenticate themselves;
- **Circumvention**: the system should be robust to malicious identification attempts.

This classification is reproduced in Table 2.1, where each trait is evaluated with respect to each of these qualities using 3 possible qualifiers: H (high), M (medium), L (low).

Biometric traits can be split in three categories:

- **physiological**, whose characteristics depend on a particular (usually static) property of the human body; examples of such traits are face, fingerprint, iris, heart sounds;

- **behavioural**, whose characteristics depend on how a person behaves; examples are gait and signature;

- **hybrid**, that contain both physiological and behavioural elements; the only trait that can be classified as hybrid is voice.

In the rest of this section, we will present a brief overview of some of the most important biometric traits.

2.3.1 Face

Face recognition is probably the identity verification method that we humans are more naturally trained to use. The face characteristics that are used for recognition are the position and shape of individual elements of the face, such as the eyes, the nose, the lips, but also the overall face shape and the relations between the elements [7].

One of the main strengths of face recognition is the simplicity of the acquisition process itself, that can be also carried covertly via surveillance cameras. Some of the most important algorithms used for feature extraction in face recognition systems are the following: Principal Components Analysis (PCA) or *eigenfaces* is an algorithm used for the representation of a face as a linear combination of base faces; Linear Discriminant Analysis (LDA) or *fisherfaces* is an algorithm that finds optimal projection vectors in the face space that maximize the ratio of intra-class dispersion to inter-class dispersion; is a graph-based algorithm, that identifies some key points in the face and creates a graph using those points and nodes, that is further processed by computing a set of complex wavelet Gabor coefficients on them.

Face recognition is still subject to noise in the acquisition phase. Illumination changes, occlusions, distance and low resolution can all affect negatively the face recognition process. Only recently we are seeing deployment of face recognition technology on consumer devices like smartphones [8], and recent advances in face recognition are helping the industry to overcome these technical difficulties and make this technology available to the mass market.

2.3.2 Voice

Along with face, voice is also a very natural means for identifying people. We are used to discriminate people by their voice, both live or on the phone.

Speaker recognition techniques can be split in two categories: text-dependent and text-independent. Text-dependent speaker recognition exploits the knowledge of what the speaker has uttered, and generally uses Hidden Markov Models (HMM) to carry on the recognition process. Text-independent speaker recognition, on the other side, do not have this knowledge, and therefore only uses the speech input data during the recognition process.

Feature extraction on speech can be done on different levels [9]. The most commonly used features are short-term spectral (or cepstral) and voice source features, like Mel-Frequency Cepstrum Coefficients (MFCC), Linear Prediction Coefficients (LPC) and Perceptual Linear Prediction (PLP). These are low-level features, that usually model more the physiological characteristics than the behavioural ones. On a higher level, there are features related to the speaking process of the person, like pitch, rhythm and temporal features; finally, there are higher-level features like accent and pronunciation. Higher level features are correlated with the behavioural component of voice biometrics; they are more robust to channel effects and noise but also significantly more difficult to extract and process.

Text-independent speaker recognition uses, among other techniques, Gaussian Mixture Models (described in Appendix A) and Support Vector Machines during the matching phase.

Voice biometrics has a wide variety of applications, including phone-based recognition and forensic biometric recognition, the latter being the context in which we will mostly discuss it in this thesis.

2.3.3 Signature

Signing a document means writing one's own full name in a designated portion of the document itself, in the act of deeming it legit, agreeing with its content and personally subscribing it. This is a practice very common today, and the authenticity of the signature has legal value.

From the biometric point of view, there are two approaches to signature recognition: using static information (off-line) [10] or using information coming from the dynamics of writing the signature (on-line) [11].

The first technique treats the signature as an image, while the second one requires the usage of devices like Wacom tablets during the acquisition phase, because they give a multi-dimensional time-varying output that usually contains, for each discrete time value, at least the position of the pen (x, y); more advanced acquisition devices also allow to capture the pressure, the azimuth and other features. From these features, a richer derived feature set can be computed (e.g., including instantaneous speed and acceleration).

On-line verification algorithms usually adopt HMMs for the modeling and matching phases.

2.3.4 Fingerprint

Fingerprints have been used for recognition for decades, especially in the forensic context. The fingerprint is the representation of the epidermis of the hand fingers, and its shape is affected both by the DNA and by the foetus grow process, so that even identical twins have different fingerprints.

A fingerprint is composed by ridges and valley, that usually have a shape similar to a concentric oval.

Most of the existing systems use as features the so-called *minutiae*, that are singularities in the ridge patterns, more specifically where ridges end and where ridges fork. There are also second-level features, like pores inside ridges, but those need higher resolution acquisition devices.

The accuracy of existing fingerprint systems is very high, and they are used in many contexts, like consumer devices authentication (e.g., laptops with built-in fingerprint scanners) and forensic contexts. The downsides of this technique are its high computational cost in large-scale identification tasks and the fact that some people do not have fingerprints suitable for recognition.

2.3.5 Iris

Iris recognition is based on the analysis of the texture (not the color) of the iris, the annular region of the eye between the pupil and the sclera. This texture is stabilized roughly after two years of life.

The techniques for iris recognition are quite mature, and there are applications in many fields, including airport border control and identification of people during war. It is a biometric modality that requires considerable co-operation from the user and has a high false rejection rate, characteristics that make it not ideal for consumer applications.

2.4 Multi-biometric systems

Biometric systems are technologically mature, and deployed in lots of contexts, from consumer hardware to military-grade access control systems. Unfortunately, This does not mean that they are flawless; the reason behind the huge amount of research in this field is the fact that there are many problems that still need to be addressed, among which the most important are [12]:

- Low robustness against noise in the biometric samples
- Non-universality
- Upper bound on matching performance
- Spoof attacks

In addition to limitations of individual traits, there are also context-dependent constraints; in web-based authentication, the problem tackled in Chapter 5, the system designer does not have any control on the acquisition devices of the user - this means that consumer-grade webcams and microphones might be used, limiting the effectiveness of each trait due to noise.

A possible solution to all these problems comes from multibiometrics [5]; multibiometrics can be defined as the usage of *multiple* techniques or sources of information at any stage of a biometric system.

There are different areas of research in the field of multibiometrics; the most important are:

- identifying the sources for multiple biometric information;

- determining the type of information to be fused;
- designing, evaluating and comparing fusion methodologies;
- building robust multimodal interfaces

Most of these ideas can be seen as application of information fusion techniques in the field of biometrics. The benefits of multibiometric systems over unibiometric systems are the following:

- improvement in the matching performance;
- increased robustness of the system to the lack of universality of the traits used;
- increased robustness of the system to spoof attacks;
- increased reliability of the system in relation to failures of individual unibiometric subsystems

The design of a multibiometric system can involve fusion on each of its components. For instance, one could design a system with two sensors for the same biometric, with one that provides actual biometric data and the other that gives some sort of liveness measurement to avoid spoofs; another example is the usage of two different biometric traits, duplicating the early stages of the individual systems and then doing a fusion of the processed data in one of the latter stages (like at score level, or at decision level).

In this thesis, we will use the multibiometric paradigm in two circumstances. The first one is the feature-level fusion of two feature sets for heart sounds, described in Section 3.3.2; the second and more obvious one is the design of a multi-modal web-based biometric authentication system, that is described in Chapter 5.

Biometric identifier	Universality	Distinctiveness	Permanence	Collectability	Performance	Acceptability	Circumvention
DNA	H	H	H	L	H	L	L
Ear	M	M	H	M	M	H	M
Face	H	L	M	H	L	H	H
Facial thermogram	H	H	L	H	M	H	L
Fingerprint	M	H	H	M	H	M	M
Gait	M	L	L	H	L	H	M
Hand geometry	M	M	M	H	M	M	M
Hand vein	M	M	M	M	M	M	L
Iris	H	H	H	M	H	L	L
Keystroke	L	L	L	M	L	M	M
Odor	H	H	H	L	L	M	L
Palmprint	M	H	H	M	H	M	M
Retina	H	H	M	L	H	L	L
Signature	L	L	L	H	L	H	H
Voice	M	L	M	L	L	M	H

Table 2.1: Comparison between biometric traits, from [1]

CHAPTER

THREE

HEART SOUNDS BIOMETRY

3.1 Introduction

Since the strength of biometric systems are highly dependent on the properties of the traits that it exploits, there are lots of research efforts towards the development of techniques that use novel traits.

As described in Chapter 2, many parts of the human body can already be used for the identification process [4]: eyes (iris and retina), face, hand (shape, veins, palmprint, fingerprints), ears, teeth etc.

The need for novel traits is motivated by the fact that there is no biometric trait that can be successfully used in all possible scenarios. So researchers are often trying to make viable the usage of traits that are complementary to the existing ones or that can act as drop-ins because they are overall better.

In this chapter, we will focus on the usage for biometric recognition purposes of an organ that is of fundamental importance for our life: the heart.

The heart is involved in the production of two biological signals, the Electrocardiograph (ECG) and the Phonocardiogram (PCG). The first is a signal derived from the electrical activity that drives the organ, while the latter is a recording of the sounds that are produced during its activity (heart sounds).

While both signals have been used as biometric traits (see [13] for ECG-based biometry), this chapter will focus on hearts-sounds biometry[1].

Using heart sounds for biometric recognition is both interesting and challenging. The main use cases for heart-sounds are quite different from the ones of the most conventional biometric traits: a consumer authentication system would rather not employ a biometric device based on heart sounds, since the acquisition is still not as easy as it is for other traits like fingerprint and face; rather, we foresee that heart sounds could be successfully used as a supplementary biometric index for high-security multi-modal identity verification system, or in critical systems based on continuous authentication.

As stated later in the chapter, it is still a novel biometric trait, and researchers still need to address many issues before the adoption of the trait will be commercially viable.

This chapter is structured as follows: in Section 3.2 we explore in-depth the comparison between heart sounds and the other conventional biometric traits; in Section 3.3 we discuss the heart sounds themselves, how they are produced by the human body, their structure and some of the algorithms that can be used for processing them; in Section 3.4 we present a review of the most important research papers that have been published in the last years on this topic; in Sections 3.5 and 3.6 we present two identity verification systems based on heart sounds, and we describe in detail their structure; in Section 3.7 we discuss the performance evaluation of the two systems, including

[1] this chapter is mainly based on the research described in [14]

the database and the evaluation protocol that have been adopted; finally, in Section 3.8 we present our conclusions and we discuss some possible topics for future research.

3.2 Comparison to other biometric traits

In Section 2.3 we presented a comparison of the most important biometric traits, as presented in [1]; in this section we will describe the relationship between those traits and heart sounds.

As a basis for the comparison, we used the 7 qualities of biometric traits depicted in Table 2.1; we added to this table a row with our subjective evaluation of heart-sounds biometry with respect to each quality, in order to give to the reader a schematic overview of the comparison. The updated table is reproduced in Table 3.1.

The reasoning behind each of our subjective evaluations of the qualities of heart sounds is as follows:

- **High Universality**: a working heart is a *conditio sine qua non* for human life;

- **Medium Distinctiveness**: the actual systems' performance is still far from the one of systems that use the most discriminating traits, and the tests are conducted using small databases; the discriminative power of heart sounds still must be demonstrated;

- **Low Permanence**: although to the best of our knowledge no studies have been conducted in this field, we perceive that heart sounds can change their properties over time, so their accuracy over extended time spans must be evaluated;

- **Low Collectability**: the collection of heart sounds is not an immediate process, and electronic stethoscopes must be placed in well-defined positions on the chest to acquire a high-quality signal;

- **Low Performance**: most of the techniques used for heart-sounds biometry are computationally intensive and, as said before, the accuracy still needs to be improved;

- **Medium Acceptability**: heart sounds can probably be perceived as unique and trustable, but people might be unwilling to use them in daily authentication tasks;

- **Low Circumvention**: it is very difficult to reproduce the heart sound of another person, and it is also difficult to record it covertly in order to reproduce it later.

The main advantages of heart sounds are, so far, the High Universality and the Low Circumvention.

The first point is undeniable and objectively true. If our body does not produce the heart sound, it means that we are not alive and so any task of authentication or live verification would be possible. This property is shared with all the biometric traits that depend on organs whose functioning is critical for our life, like the brain. This also means that heart sounds cannot be used if the subject is not close to the sensor when the signal needs to be recorded, making it useless for situation like crime scene analysis.

The second point is maybe less undeniable but still true. As recording heart sounds is more difficult than recording - for instance - voice, so a malicious user of the system would have trouble in recording another user's heart sounds for using them later, not mentioning the difficulties in hiding a proper

Biometric identifier	Universality	Distinctiveness	Permanence	Collectability	Performance	Acceptability	Circumvention
DNA	H	H	H	L	H	L	L
Ear	M	M	H	M	M	H	M
Face	H	L	M	H	L	H	H
Facial thermogram	H	H	L	H	M	H	L
Fingerprint	M	H	H	M	H	M	M
Gait	M	L	L	H	L	H	M
Hand geometry	M	M	M	H	M	M	M
Hand vein	M	M	M	M	M	M	L
Iris	H	H	H	M	H	L	L
Keystroke	L	L	L	M	L	M	M
Odor	H	H	H	L	L	M	L
Palmprint	M	H	H	M	H	M	M
Retina	H	H	M	L	H	L	L
Signature	L	L	L	H	L	H	H
Voice	M	L	M	L	L	M	H
Heart sounds	H	M	L	L	L	M	L

Table 3.1: Relation between heart sounds and other biometric traits

audio device that plays it back if the acquisition phase is done under supervision. One possibility that must still be explored is the usage of synthetic heart sounds; since most of the approaches today use generative models (like GMMs), the attacker could steal the templates and try to use them to generate fake heart sounds that share the biometric properties of the user to which the template belongs to. This concerns, however, are more likely to be addressed by researchers in the area of the security of biometric systems. Our studies are more focused on the investigation of the biometric properties of heart sounds, and therefore ignore these problems, because they usually come into play at a later stage of research, when the trait has reached sufficient maturity.

The main drawbacks of heart-sounds biometry are probably the Low Performance and, above all, its overall immaturity as a biometric trait. Of course, heart-sounds biometry is a new technique, and as such many of its current drawbacks will probably be addressed and resolved in future research work.

3.3 Heart sounds

3.3.1 Physiology and structure of heart sounds

The heart sound signal is a complex, non-stationary and quasi-periodic signal that is produced by the heart during its continuous pumping work [15]. It is composed by several sounds, each associated with a specific event in the working cycle of the heart.

Heart sounds fall in two categories:

- **primary sounds**, produced by the closure of the heart valves;
- **other sounds**, produced by the blood flowing in the heart or by pathologies;

The primary sounds are S1 and S2. The first sound, S1, is caused by the closure of the tricuspid and mitral valves, while the second sound, S2, is caused by the closure of the aortic and pulmonary valves.

Among the other sounds, there are the S3 and S4 sounds, that are quieter and less frequent than S1 and S2, and murmurs, that are high-frequency noises.

Most of these smaller sounds are periodic, and their frequency is usually measured in beats per minute (bpm). In each heart beat there are one S1 and one S2 sound. We refer to each heart beat as a “cardiac cycle”.

In our systems, we only use the primary sounds because they are the two loudest sounds and they are the only ones that a heart always produces, even in pathological conditions. We separate them from the rest of the heart sound signal using the algorithm described in Section 3.3.2.

3.3.2 Processing heart sounds

Heart sounds are monodimensional signals, and can be processed, to some extent, using techniques known to work on other monodimensional signals, like audio signals. Those techniques then need to be refined taking into account the peculiarities of the signal, its structure and components.

Once the heart signal is acquired, we need to execute three kinds of task with it:

- pre-processing
- segmentation
- feature extraction

The rest of the biometric systems do not deal with the signal itself, but with features or templates (models); in this section we will instead describe the algorithms that operate on the heart sound signal.

Pre-processing is carried out with standard signal processing techniques, like low-pass filtering, that does not need to be discussed.

Segmentation is the task of separating the S1 and S2 sounds from the rest of the signal, and is discussed in Section 3.3.2.

Feature extraction, as discussed in Section 2.1 is the task of transforming the signal into an alternate, more compact and possibly more meaningful representation. We present three algorithms for feature extraction, two that work in the frequency domain (CZT and MFCC) and one that works in the time domain (FSR).

Segmentation

In this section we describe a variation of the algorithm described in [16] to separate the S1 and S2 tones from the rest of the heart sound signal, improved to deal with long heart sounds.

Such a separation is necessary because we believe that the S1 and S2 tones are as important to heart sounds as the vowels are to the voice signal. They can be considered stationary in the short term and they convey significant biometric information, that is then processed by feature extraction algorithms.

A simple energy-based approach can not be used because the signal can contain impulsive noise that could be mistaken for a significant sound.

Before being processed by the algorithm, the signal is split in frames. Usually, 20ms wide frames are used, with 10ms overlap between frames.

The first step of the algorithm is searching the frame with the highest energy, that is called SX1. At this stage, we do not know if we found an S1 or an S2 sound.

Then, in order to estimate the frequency of the heart beat, and therefore the period P of the signal, the maximum value of the autocorrelation function is computed. Low-frequency components are ignored by searching only over the portion of autocorrelation after the first minimum.

The algorithm then searches other maxima to the left and to the right of SX1, moving by a number P of frames in each direction and searching for local maxima in a window of the energy signal in order to take into account small fluctuations of the heart rate. After each maximum is selected, a constant-width window is applied to select a portion of the signal.

After having completed the search that starts from SX1, all the corresponding frames in the original signal are zeroed out, and the procedure is repeated to find a new maximum-energy frame, called SX2, and the other peaks are found in the same way.

Finally, the positions of SX1 and SX2 are compared, and the algorithm then decides if SX1, and all the frames found at distance P starting from it, must be classified as S1 or S2; the remaining identified frames are classified accordingly.

The nature of this algorithm requires that it works only on short sequences, 4 to 6 seconds long, because as the sequence gets longer the periodicity of the sequence fades away due to noise and variations of the heart rate.

To overcome this problem, the signal is split into 4-seconds wide windows and the algorithm is applied to each window. The resulting sets of heart sounds endpoint are then joined into a single set.

The chirp z-transform

The Chirp z-Transform (CZT) is an algorithm for the computation of the z-Transform of sampled signals that offers some additional flexibility with re-

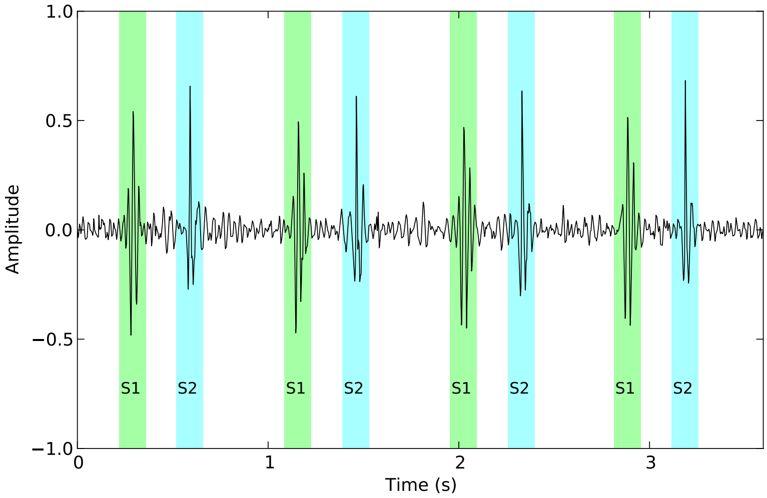


Figure 3.1: Example of S1 and S2 detection

spect to the Fast Fourier Transform (FFT) algorithm.

The main advantage of the CZT exploited in the analysis of heart sounds is the fact that it allows high-resolution analysis of narrow frequency bands, offering higher resolution than the FFT.

For more details on the CZT, please refer to [17]

Cepstral analysis

Mel-Frequency Cepstrum Coefficients (MFCC) are one of the most widespread parametric representation of audio signals [18].

The basic idea of MFCC is the extraction of cepstrum coefficients using a non-linearly spaced filterbank; the filterbank is instead spaced according to the Mel Scale: filters are linearly spaced up to 1 kHz, and then are logarithmically spaced, decreasing detail as the frequency increases. Parametric representation that use only linearly spaced filters are called Linear Frequency Cepstrum Coefficients (LFCC) or Linear Frequency Bands Cepstra (LFBC).

This scale is useful because it takes into account the way we perceive sounds.

The relation between the Mel frequency $\hat{f}_{mel}$ and the linear frequency f_{lin} is the following:

$$\hat{f}_{mel} = 2595 \cdot \log_{10}\left(\frac{1 + f_{lin}}{700}\right) \tag{3.1}$$

The first step of the algorithm is to compute the FFT of the input signal; the spectrum is then fed to the filterbank, and the *i*-th cepstrum coefficient is computed using the following formula:

$$C_i = \sum_{k=1}^{K} X_k \cdot \cos\left(i \cdot \left(k - \frac{1}{2}\right) \cdot \frac{\pi}{K}\right) i = 0, \ldots, M \tag{3.2}$$

Where K is the number of filters in the filterbank, X_k is the log-energy output of the *k*-th filter and M is the number of coefficients that must be computed.

Many parameters have to be chosen when computing cepstrum coefficients. Among them: the bandwidth and the scale of the filterbank (Mel vs. linear), the number and spectral width of filters, the number of coefficients.

In addition to this, differential cepstrum coefficients, typically denoted using a Δ (first order) or $\Delta\Delta$ (second order), can be computed and used.

Figure 3.2 shows an example of three S1 sounds and the relative MFCC spectrograms; the first two (a, b) belong to the same person, while the third (c) belongs to a different person.

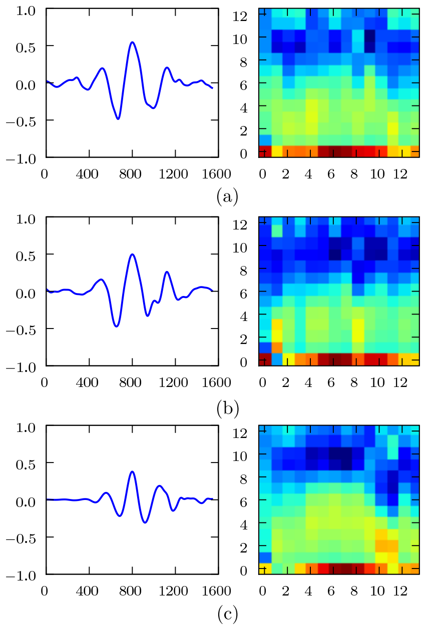


Figure 3.2: Example of waveforms and MFCC spectrograms of S1 sounds

The First-to-Second Ratio (FSR)

In addition to standard feature extraction techniques, it is desirable to develop ad-hoc features for the heart sound, as it is not a simple audio sequence but has specific properties that could be exploited to develop features with additional discriminative power.

This is why in [19] we introduced a time-domain feature called First-to-Second Ratio (FSR). Intuitively, the FSR represents the power ratio of the first heart sound (S1) to the second heart sound (S2). During our experiments, we observed that some people tend to have an S1 sound that is louder than S2, while in others this relation is inverted. We try to represent this diversity using our new feature.

The implementation of the feature is different in the two biometric systems that we will describe in this chapter, and a discussion of the two algorithms can be found in later sections.

Figure 3.3 shows the plot of the distribution of the intra-person and inter-person d_{FSR} distances, as defined in Equation 3.7. The comparison was carried over a database of 50 people. This plot clearly shows that the FSR has some discriminative power.

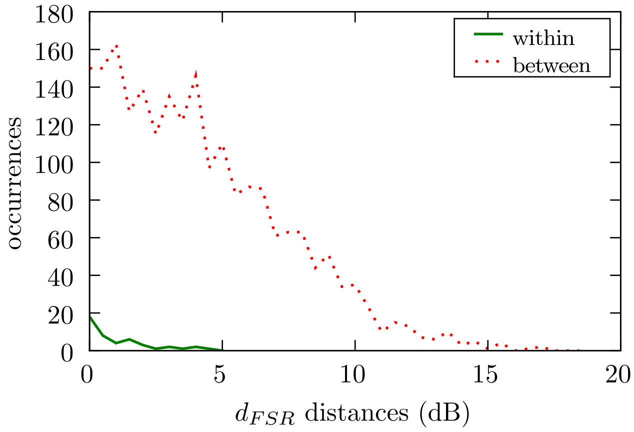


Figure 3.3: Distribution of intra-person and inter-person FSR distances

3.4 Review of related works

In the last years, different research groups have been studying the possibility of using heart sounds for biometric recognition. In this section, we will briefly describe their methods.

In Table 3.2 we summarized the main characteristics of the works that will be analyzed in this section, using the following criteria:

- **Database** - the number of people involved in the study and the amount of heart sounds recorded from each of them;

- **Features** - which features were extracted from the signal, at frame level or from the whole sequence;

- **Classification** - how features were used to make a decision.

We chose not to represent performance in this table for two reasons: first, most papers do not adopt the same performance metric, so it would be difficult to compare them; second, the database and the approach used are quite different one from another, so it would not be a fair comparison.

In the rest of the section, we will briefly review each of these papers.

[20] was one of the first works in the field of heart-sounds biometry. In this paper, the authors first do a quick exploration of the feasibility of using heart sounds as a biometric trait, by recording a test database composed of 128 people, using 1-minute heart sounds and splitting the same signal into a training and a testing sequence. Having obtained good recognition performance using the HTK Speech Recognition toolkit, they do a deeper test using a database recorded from 10 people and containing 100 sounds for each person, investigating the performance of the system using different feature extraction algorithms (MFCC, LFBC), different classification schemes (Vector Quantization (VQ) and Gaussian Mixture Models (GMM)) and investigating the impact of the frame size and of the training/test length. After testing many combinations of those parameters, they conclude that, on their database, the most performing system is composed of LFBC features (60 cepstra + log-

Paper	Database	Features	Classification
[20]	10 people 100 HS each	MFCC LBFC	GMM VQ
[21]	52 people 100m each	Multiple	SVM
[22]	10 people 20 HS each	Energy peaks	Euclidean distance
[23]	21 people 6 HS each 8 seconds per HS	MFCC, LDA, energy peaks	Euclidean distance
[24]	40 people 10 HS 10 seconds per HS	autocorrelation cross-correlation complex cepstrum	MSE kNN

Table 3.2: Comparison of recent works about heart-sound biometrics

energy + 256ms frames with no overlap), GMM-4 classification, 30s of training/test length.

The authors of [21], one of which worked on [20], take the idea of finding a good and representative feature set for heart sounds even further, exploring 7 sets of features: temporal shape, spectral shape, cepstral coefficients, harmonic features, rhythmic features, cardiac features and the GMM supervector. They then feed all those features to a feature selection method called RFE-SVM and use two feature selection strategies (optimal and sub-optimal) to find the best set of features among the ones they considered. The tests were conducted on a database of 52 people and the results, expressed in terms of Equal Error Rate (EER), are better for the automatically selected feature sets with respect to the EERs computed over each individual feature set.

In [22], the authors describe an experimental system where the signal is first downsampled from 11025 Hz to 2205 Hz; then it is processed using the Discrete Wavelet Transform, using the Daubechies-6 wavelet, and the D4 and D5 sub-bands (34 to 138 Hz) are then selected for further processing. After a normalization and framing step, the authors then extract from the signal some energy parameters, and they find that, among the ones considered, the Shannon energy envelogram is the feature that gives the best performance on their database of 10 people.

The authors of [23] do not propose a pure-PCG approach, but they rather investigate the usage of both the ECG and PCG for biometric recognition. In this short summary, we will focus only on the part of their work that is related to PCG. The heart sounds are processed using the Daubechies-5 wavelet, up to the 5th scale, and retaining only coefficients from the 3rd, 4th and 5th scales. They then use two energy thresholds (low and high), to select which coefficients should be used for further stages. The remaining frames are then processed using the Short-Term Fourier Transform (STFT), the Mel-Frequency filterbank and Linear Discriminant Analysis (LDA) for dimensionality reduction. The decision is made using the Euclidean distance from the feature vector obtained in this way and the template stored in the database. They test the PCG-based system on a database of 21 people, and their combined PCG-ECG systems has better performance.

The authors of [24] filter the signal using the DWT; then they extract different kinds of features: auto-correlation, cross-correlation and cepstra. They then test the identities of people in their database, that is composed by 40 people, using two classifiers: Mean Square Error (MSE) and k-Nearest Neighbor (kNN). On their database, the kNN classifier performs better than the MSE one.

3.5 The structural approach to heart-sounds biometry

The first system that we describe in depth was introduced in [16]; it was designed to work with short heart sounds, 4 to 6 seconds long and thus containing at least four cardiac cycles (S1-S2).

The restriction on the length of the heart sound was removed in [25], that introduced the quality-based best subsequence selection algorithm, described in 3.5.1.

We call this system "structural" because the identity templates are stored as feature vectors, in opposition to the "statistical" approach, that does not directly keep the feature vectors but instead it represents identities via statistical parameters inferred in the learning phase.

Figure 3.4 contains the block diagram of the system. Each of the steps will be described in the following sections.

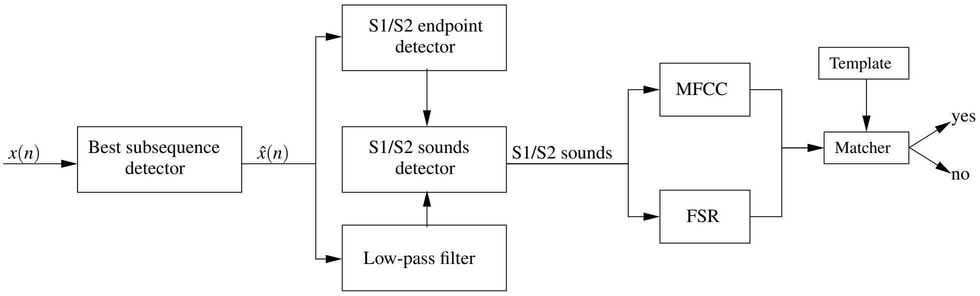


Figure 3.4: Block diagram of the proposed cardiac biometry system

3.5.1 The best subsequence selection algorithm

The fact that the segmentation and matching algorithms of the original system were designed to work on short sequences was a strong constraint for the system. It was required that a human operator selected a portion of the input signal based on some subjective assumptions. It was clearly a flaw that needed to be addressed in further versions of the system.

To resolve this issue, the authors developed a quality-based subsequence selection algorithm, based on the definition of a quality index $DHS_{QI}(i)$ for each contiguous subsequence i of the input signal.

The quality index is based on a cepstral similarity criterion: the selected subsequence is the one for which the cepstral distance of the tones is the lowest possible. So, for a given subsequence i, the quality index is defined as:

$$DHS_{QI}(i) = \frac{1}{\sum_{k=1}^{4} \sum_{\substack{j=1 \\ j \neq k}}^{4} d_{S1}(j,k) + \sum_{k=1}^{4} \sum_{\substack{j=1 \\ j \neq k}}^{4} d_{S2}(j,k)} \tag{3.3}$$

Where d_{S1} and d_{S2} are the cepstral distances defined in 3.5.5.

The subsequence $\bar{i}$ with the maximum value of $DHS_{QI}(\bar{i})$ is then selected as the best one and retained for further processing, while the rest of the input signal is discarded.

3.5.2 Filtering and segmentation

After the best subsequence selection, the signal is then given in input to the heart sound endpoint detection algorithm described in 3.3.2.

The endpoints that it finds are then used to extract the relevant portions of the signal over a version of the heart sound signal that was previously filtered using a low-pass filter, which removes the high-frequency extraneous components.

3.5.3 Feature extraction

The heart sounds are then passed to the feature extraction module, that computes the cepstral features according to the algorithm described in 3.3.2.

This system uses $M = 12$ MFCC coefficients, with the addition of a 13-th coefficient computed using an $i = 0$ value in Equation 3.2, that is the log-energy of the analyzed sound.

3.5.4 Computation of the First-to-Second Ratio

For each input signal, the system computes the FSR according to the following algorithm.

Let N be the number of complete S1-S2 cardiac cycles in the signal. Let P_{S1_i} (resp. P_{S2_i}) be the power of the i-th S1 (resp. S2) sound.

We can then define $\overline{P_{S1}}$ and $\overline{P_{S2}}$, the average powers of S1 and S2 heart sounds:

$$\overline{P_{S1}} = \frac{1}{N}\sum_{i=1}^{N} P_{S1_i} \tag{3.4}$$

$$\overline{P_{S2}} = \frac{1}{N}\sum_{i=1}^{N} P_{S2_i} \tag{3.5}$$

Using these definitions, we can then define the First-to-Second Ratio of a given heart sound signal as:

$$FSR = \frac{\overline{P_{S1}}}{\overline{P_{S2}}} \tag{3.6}$$

For two given heart sounds x_1 and x_2, we define the FSR distance as:

$$d_{FSR}(x_1, x_2) = |FSR_{dB}(x_1) - FSR_{dB}(x_2)| \tag{3.7}$$

3.5.5 Matching and identity verification

The crucial point of identity verification is the computation of the distance between the feature set that represents the input signal and the template associated with the identity claimed in the acquisition phase by the person that is trying to be authenticated by the system.

This system employs two kinds of distance: the first in the cepstral domain and the second using the FSR.

MFCC are compared using the Euclidean metric (d_2). Given two heart sound signals X and Y, let $X_{S1}(i)$ (resp. $X_{S2}(i)$) be the feature vector for the i-th S1 (resp. S2) sound of the X signal and Y_{S1} and Y_{S2} the analogous vectors for the Y signal. Then the cepstral distances between X and Y can be defined as follows:

$$d_{S1}(X,Y) = \frac{1}{N^2} \sum_{i,j=1}^{N} d_2(X_{S1}(i), Y_{S1}(j)) \tag{3.8}$$

$$d_{S2}(X,Y) = \frac{1}{N^2} \sum_{i,j=1}^{N} d_2(X_{S2}(i), Y_{S2}(j)) \tag{3.9}$$

Now let us take into account the FSR. Starting from the d_{FSR} as defined in Equation 3.7, we wanted this distance to act like an amplifying factor for the cepstral distance, making the distance bigger when it has an high value while not changing the distance for low values.

We then normalized the values of d_{FSR} between 0 and 1 ($d_{FSR_{norm}}$), we chose a threshold of activation of the FSR (th_{FSR}) and we defined k_{FSR}, an amplifying factor used in the matching phase, as follows:

$$k_{FSR} = \max\left(1, \frac{d_{FSR_{norm}}}{th_{FSR}}\right) \tag{3.10}$$

In this way, if the normalized FSR distance is lower than th_{FSR} it has no effect on the final score, but if it is larger, it will increase the cepstral distance.

Finally, the distance between X and Y can be computed as follows:

$$d(X,Y) = k_{FSR} \cdot \sqrt{d_{S1}(X,Y)^2 + d_{S2}(X,Y)^2} \tag{3.11}$$

3.6 The statistical approach to heart-sounds biometry

In opposition to the system analyzed in Section 3.5, the one that will be described in this section is based on a machine learning process that does not directly compare the features extracted from the heart sounds, but instead uses them to infer a statistical model of the identity and makes a decision computing the probability that the input signal belongs to the person whose identity was claimed in the identity verification process.

Figure 3.5 contains the block diagram of the system. Each of the steps will be described in the following sections.

This system uses the GMM-UBM approach for the creation of the identity models and for the computation of the scores, in the form of log-likelihood ratios. For more information about the GMM-UBM method, see Appendix A.

3.6.1 Front-end processing

Each time the system gets an input file, whether for training a model or for identity verification, it goes through some common steps.

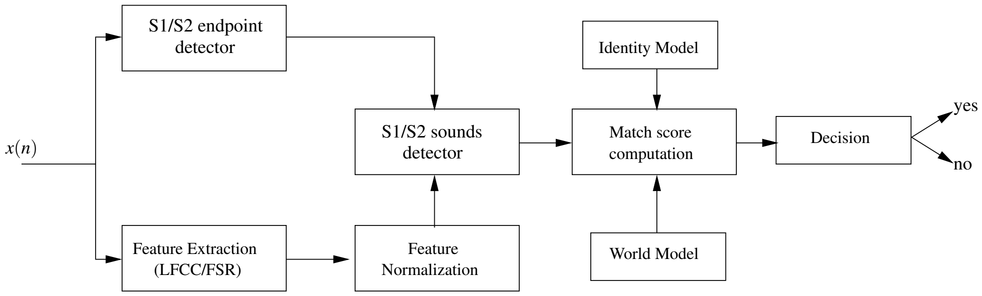


Figure 3.5: Block diagram of the Statistical system

First, heart sounds segmentation is carried on, using the algorithm described in Section 3.3.2.

Then, cepstral features are extracted using a tool called *sfbcep*, part of the SPro suite [26].

For each input sequence, we also compute the FSR and we append it to each feature vector, as a sequence-wise feature. This operation is done by a program developed using the low-level Alize framework.

In the context of the statistical approach, it seemed more appropriate to just append the FSR to the feature vector computed from each frame in the feature extraction phase, and then let the GMM algorithms generalize this knowledge.

To do this, we split the input heart sound signal in 5-second windows and we compute an average FSR ($\overline{FSR}$) for each signal window, that is the average of the FSR values of each heart cycle.

3.6.2 The experimental framework

The experimental set-up created for the evaluation of this technique was implemented using some tools provided by ALIZE/SpkDet , an open source toolkit for speaker recognition described in Appendix B.

The adaptation of parts of a system designed for speaker recognition to a different problem was possible because the toolkit is sufficiently general and flexible, and because the features used for heart-sounds biometry are similar to the ones used for speaker recognition, as outlined in Section 3.3.2.

During the world training phase, the system estimates the parameters of the world model λ_W using a randomly selected subset of the input signals.

The identity models λ_i are then derived from the world model W using the Maximum A-Posteriori (MAP) algorithm.

During identity verification, the matching score is computed using Equation A.3, and the final decision is taken comparing the score to a threshold (θ).

3.6.3 Optimization of the method

During the development of the system, some parameters have been tuned in order to get the best performance. Namely, three different cepstral feature sets have been considered in [27]:

- $16 + 16\,\Delta + E + \Delta E$
- $16 + 16\,\Delta + 16\,\Delta\Delta$
- $19 + 19\,\Delta + E + \Delta E$

However, the first of these sets proved to be the most effective

In [28] the impact of the FSR and of the number of Gaussian densities in the mixtures was studied. Four different model sizes (128, 256, 512, 1024) were tested, with and without FSR, and the best combination of those parameters, on our database, is 256 Gaussians with FSR.

3.7 Performance evaluation

In this section, we will compare the performance of the two systems described in the previous sections of this chapter. We will first describe the database in Section 3.7.1, then the performance evaluation protocol in 3.7.2. Finally the results of the comparison will be given in Section 3.7.3.

3.7.1 The HSCT-11 database

One of the drawbacks that is relatively common among novel biometric traits is the absence of significantly large databases for performance evaluation.

To overcome this problem, we are building a heart sounds database suitable for identity verification performance evaluation. This database is called HSCT-11, that stands for Heart Sounds Catania 2011[2] [29].

Currently, in the database there are sounds recorded from 206 people, 157 male and 49 female; for each person, there are two separate recordings, each lasting from 20 to 70 seconds; the average length of the recordings is 45 seconds. The heart sounds have been acquired using a Thinklabs Rhythm Digital Electronic Stethoscope, connected to a computer via an audio card. The sounds have been converted to the Wave audio format, using 16 bit per second and at a rate of 11025 Hz.

The filenames encode the following metadata about the person:

- the first character encodes the sex of the person (M or F);
- the next 4 characters are the numeric unique ID of the person;

[2] the database is available for free download at the URL http://www.diit.unict.it/users/spadaccini/hsct11

- the next character encodes the heart valve used for the auscultation (M: mitral, P: pulmonary, A: aortic, T: tricuspid); this database currently contains only sequences recorded near the pulmonary valve;
- the next character encodes whether the recording was done with the subject in resting condition (N) or after some light physical activity (C); so far the database contains only sequences recorded in resting condition;
- the next 3 characters encode the sequential number of the registration acquired from a given person; the first of these 3 characters is always the letter R.
- the next 7 characters encode the date of the acquisition; the first one is always a letter D, the others represent the date in the format MMDDYY;
- the next 7 characters encode the birth date of the subject; the first one is always a letter N, the others represent the date in the format MMDDYY;

The letters between fields could have been avoided since the fields have a fixed length, but they have been inserted because they make it easier for human eyes to scan the filename and extract the required information. An example filename is: F7007NR01D290610N051077.wav.

3.7.2 Evaluation protocol

The comparison has been done in the following way: for each person, one sequence is used for the model training phase and one is used for the computation of matching scores.

Let X be a given person, X_a its first recording and X_b its second recording; also let D be the set of all the people in the database, and let $N = |D| = 206$

be the number of people in it. Let S be the matching function that, given an identity model and a recording gives a similarity score.

For each person, the database user should compute one genuine matching score, that is $S(M_X, X_b)$, and $N-1$ impostor matching scores $S(M_Y, X_b), \forall Y \in \{D \setminus X\}$. This will yield N genuine matching scores and $N \cdot (N-1)$ impostor matching scores.

3.7.3 Results

The performance of our two systems has been computed over the HSCT-11 database, and the results are reported in Table 3.3.

System	**EER (%)**
Structural	36.86
Statistical	**13.66**

Table 3.3: EER of the two heart-sounds biometry systems

The huge difference in the performance of the two systems reflects the fact that the first one is not being actively developed since 2009, and it was designed to work on small databases, while the second has already proved to work well on larger databases.

It is important to highlight that, in spite of a 25% increment of the size of the database, the error rate remained almost constant with respect to the last evaluation of the system, in which a test over a 165 people database yielded a 13.70% EER.

Figure 3.6 shows the Detection Error Trade-off (DET) curves of the two systems. As stated before, a DET curve shows how the analyzed system per-

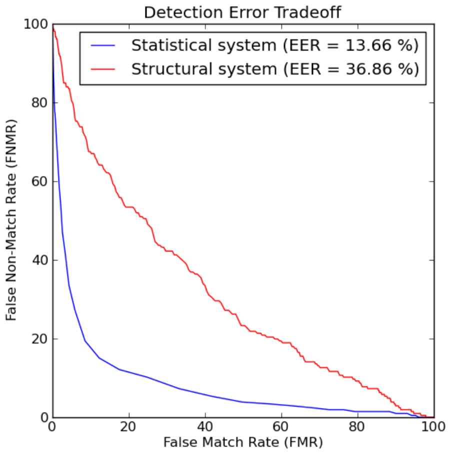


Figure 3.6: Detection Error Trade-off (DET) curves of the two systems

forms in terms of false matches/false non-matches as the system threshold is changed.

In both cases, fixing a false match (resp. false non-match) rate, the system that performs better is the one with the lowest false non-match (resp. false match) rate.

Looking at Figure 3.6, it is easy to understand that the statistical system performs better in both high-security (e.g., FMR = 1-2%) and low-security (e.g., FNMR = 1-2%) setups.

We can therefore conclude that the statistical approach is definitely more promising that the structural one, at least with the current algorithms and using the database described in 3.7.1.

3.8 Conclusions

In this chapter, we described a novel biometric technique based on heart sounds, analyzing two different approaches and evaluating their performance.

As shown in the survey of recent papers in this field, the number of research groups that work on this topic is slowly increasing, and this means that there is interest for this new biometric trait.

The performance analysis of the two systems presented in this chapter show that, using our database, the performance of the system is still not suitable for real-world systems but it is not so far from being viable.

Unfortunately, it is difficult to compare the results with the ones presented in the other papers, because the databases are not public, the performance metrics are often different from the EER and the size of the databases is just too different to try and derive any conclusion from comparing the numeric values obtained from the evaluation.

To the best of our knowledge, our database is so far the one that contains the highest number of people, and while this is encouraging because it might give more validity to our results, we recognize that 206 people are still not enough to draw significant statistical conclusions. Moreover, the database needs to be more diversified, with more than 2 sessions and with more variability in session time spread, health condition etc.

As larger databases of heart sounds become available to the scientific community, there are some issues that need to be addressed in future research.

First of all, the identification performance should be kept low even for larger databases. This means that the matching algorithms will be fine-tuned and a suitable feature set will be identified, probably containing both elements from the frequency domain and the time domain.

Next, the mid-term and long-term reliability of heart sounds will be assessed, analyzing how their biometric properties change as time goes by. Additionally, the impact of cardiac diseases on the identification performance will be assessed.

Finally, when the algorithms will be more mature and several independent scientific evaluations will have given positive feedback on the idea, some practical issues like computational efficiency will be tackled, and possibly adhoc sensors with embedded matching algorithms will be developed, thus making heart-sounds biometry a suitable alternative to the mainstream biometric traits.

CHAPTER

FOUR

FORENSIC SPEAKER RECOGNITION

4.1 Introduction

In this chapter we will analyze some of the current research problems in the field of forensic speaker verification, and we will present the results of our experiments[1].

Speaker recognition has already been described in Section 2.3.2, and it is a biometric technique that is employed in many different contexts, with various degrees of success.

In this chapter, we are interested in a narrow context: the analysis of speech data coming from wiretappings or ambient recordings retrieved during criminal investigation with the purpose of being able of recognizing if a given sentence had been uttered by a given person. This process is called forensic speaker recognition, and it is still a controversial topic [33].

[1] this chapter is based on the research described in [30–32]

As of the writing of this document, in Italian courts this process is still carried on using semi-automatic techniques. This means that an expert witness does the analysis with the aid of some specialized software, but he is free to change some parameters that can affect the final outcome of the identification.

It is obvious that human errors, or in the worst case conscious alterations of the parameters, can lead to wrong results, with disastrous consequence on the trial.

What we want to analyze is how can state-of-the-art speaker recognition techniques be employed in this context, what are their limitations and their strengths and what must be improved in order to migrate from old-school manual or semi-automatic techniques to new, reliable and objective automatic methods.

It is well-known that speech signal quality is of fundamental importance for accurate speaker identification [34].

The reliability of a speech biometry system is known to depend on the amount of material available, in particular on the number of vowels present in the sequence being analysed, and the quality of the signal [35]. The former affects the resolution power of the system, while the latter impacts the correct estimation of biometric indexes. In automatic or semi-automatic speaker recognition, background noise is one of the main causes of alteration of the acoustic indexes used in the biometric identification/verification phase [36]. Therefore, background noise is one of the main causes of a performance degradation of a biometry system.

In this chapter, we will analyze the behaviour of some speaker recognition techniques when the conditions are not controlled and the speech sequences are disturbed by background noise.

In Section 4.2 we will describe the speech and noise databases used for the experiments; in Section 4.3 we will analyze the performance of two Signal-to-Noise (SNR) estimation algorithms; in Section 4.4 we will analyze the performance of a speaker recognition toolkit; in Section 4.5 we will analyze the impact of Voice Activity Detection (VAD) algorithms on the recognition rates; finally, in Section 4.6 we will draw our conclusions.

4.2 Speech and noise databases

4.2.1 The TIMIT speech database

All the speaker recognition experiments described in this chapter use as a speech database a subset of the TIMIT (Texas Instrument Massachusetts Institute of Technology) database, that will be briefly described in this section.

The TIMIT database contains speech data acquired from 630 people, that are split in subsets according to the Dialect Region to which each of them belongs. Each DR is further split in training and test set. The number of speakers contained in each DR, and their division in training and test set are reported in Table 4.1.

This database was explicitly designed to provide speech researchers with a phonetically rich dataset to use for research in speech recognition, but it is widely adopted also in the speaker recognition research community.

It contains three types of sentences, dialectal (SA), phonetically-compact (SX) and phonetically diverse (SI). The total number of spoken sentences is 6300, 10 for each of the 630 speakers. There is some superposition between speakers, because there are sentences that are spoken by more than one person. Each person, however, has to read 2 SA sentences, 5 SX sentences and 3 SI sentences.

The database also contains annotations about the start and end points of different lexical tokens (phonemes, words and sentences). This was especially

Dialect Region	Total	Training	Test
New England (DR1)	49	38	11
Northern (DR2)	102	76	26
North Midland (DR3)	102	76	26
South Midland (DR4)	100	68	32
Southern (DR5)	98	70	28
New York City (DR6)	46	35	11
Western (DR7)	100	77	23
Moved around (DR8)	33	22	11
Totals:	630	462	168

Table 4.1: Composition of the TIMIT data set

useful for the research on SNR and VAD, because we could compare our algorithms with ground truth provided by the database itself.

4.2.2 The noise database

The noise database comprises a set of recordings of different types of background noise, each lasting 3 minutes, sampled at 8 kHz and linearly quantized using 16 bits per sample. The types of noise contained in the database fall into the following categories:

- **Car**, recordings made inside a car;
- **Office**, recordings made inside an office during working hours;
- **Factory**, recordings made inside a factory;

- **Construction**, recordings of the noise produced by the equipment used in a building site;

- **Train**, recordings made inside a train;

4.3 Performance evaluation of SNR estimation algorithms

In forensics, one of the most widely adopted methods to assess the quality of the intercepted signal is based on the estimation of the Signal to Noise Ratio (SNR), that should not be lower than a critical threshold, usually chosen between 6 and 10 dB [35]. It is possible to estimate the SNR using manual or semi-automatic methods. Both of them exploit the typical ON-OFF structure of conversations, which means that on average there are times when there is a speech activity (talkspurt) and times when nobody is talking, and the signal is mainly composed by environmental noise recorded by the microphone (background noise). With manual methods, the SNR is estimated choosing manually the segment of talkspurt and the segment of background noise immediately before or after the talkspurt. The estimation is computed by the following formula:

$$SNR_{est} = \frac{P_{talk} - P_{noise}}{P_{noise}} \tag{4.1}$$

Semi-automatic estimation methods use a Voice Activity Detection (VAD) algorithm that separates the ON segments from the OFF segments in a given conversation, and use those segments to estimate the SNR using equation 4.1 [37, 38].

Both algorithms do not give an exact value of the SNR, because the noise sampled for the SNR estimation is different from the noise that degraded the vocal segment for which the SNR is being estimated. This happens because the noise level can be measured only when the speakers are not talking, in an OFF segment.

Sometimes the estimation error causes the elimination of good-quality data (under-estimation of the SNR), while sometimes it causes the usage of low-quality biometric data that was probably corrupted by noise in the subsequent identification process (over-estimation of the SNR)

In this section, we will discuss about the accuracy of the SNR estimation methods, comparing their average estimation error to the real SNR.

4.3.1 Speech and background noise database

In this experiment, we used speech data coming from 100 people, half female and half male, randomly selected from the DR1 subset of the TIMIT database

The 10 sentences spoken by each person, sampled at 8 kHz and linearly quantized using 16 bits per sample, have been used to produce a clean conversation composed by talkspurt segments (ON) normalized to an average power level of $-26dB_{ovl}$ and silence segments (OFF). The ON-OFF statistics were chosen using the model proposed in [39].

We used 4 kinds of background noise: Car, Office, Stadium, Construction.

For each type of noise, the clean sequence was digitally summed to the noise, in order to get sequences with four different real SNRs in the activity segments: 0, 10, 20 and 30 dB.

4.3.2 SNR estimation methods

Both analyzed SNR estimation methods exploit the manual phonetic marking offered by the TIMIT database. In particular, for the sake of simplicity, we selected a restricted subset of vowel sounds (“ae”, “iy”, “eh”, “ao”), of which only the central 20 ms were considered.

The manual SNR estimation method computes the estimated SNR as the ratio between the power of the signal of the current vowel, (P_{talk}) lasting 20ms, to the power of noise, (P_{noise}), measured at the nearest OFF segment and lasting 20 ms. The classification of the signal in ON and OFF segments is done manually.

The semi-automatic method uses the VAD algorithm to automatically classify ON and OFF segments. The VAD used is the one standardized by the ETSI for the speech codec AMR [40]. In this case, the noise power is measured using the nearest OFF segment classified by the VAD.

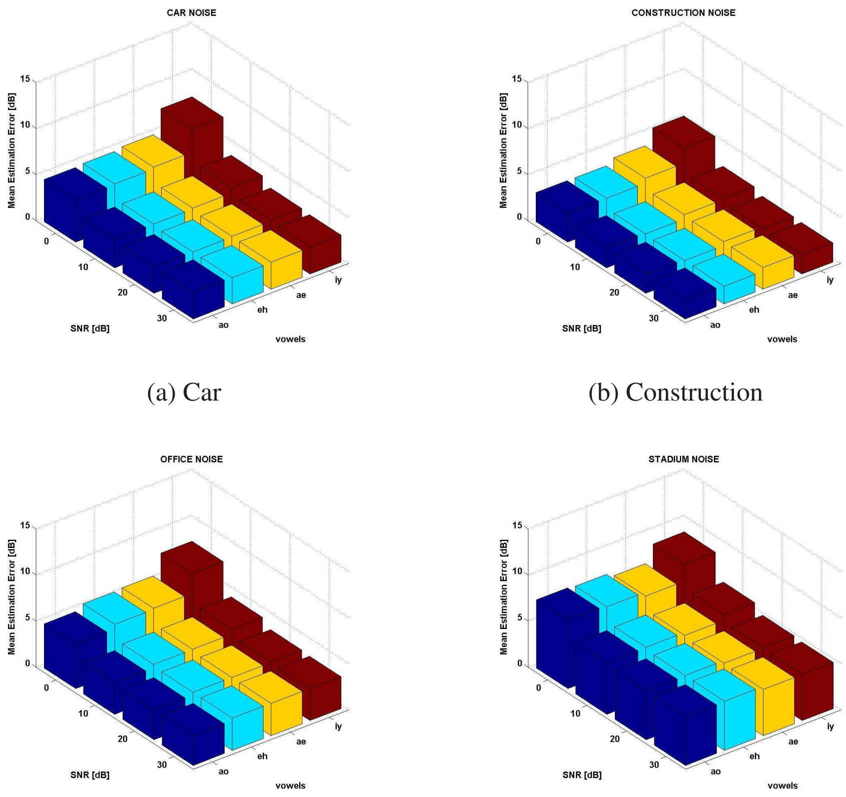
The values obtained by the two algorithms have then been compared to the real SNR, computed as the ratio between the power of the vowel measured on the clean signal and the power of the background noise measured in the same temporal position but on the noise sequence.

4.3.3 Results

The first analysis that we present is the computation of the average estimation errors. In each subplot, two axis represent the SNR and the vowel, while the third one represents the average estimation error.

Figure 4.1 shows the average estimation error for the manual method, while Figure 4.2 shows the same error, but for the semi-automatic method.

The performance of both methods are similar for the Car and Noise, with



(a) Car

(b) Construction

(c) Office

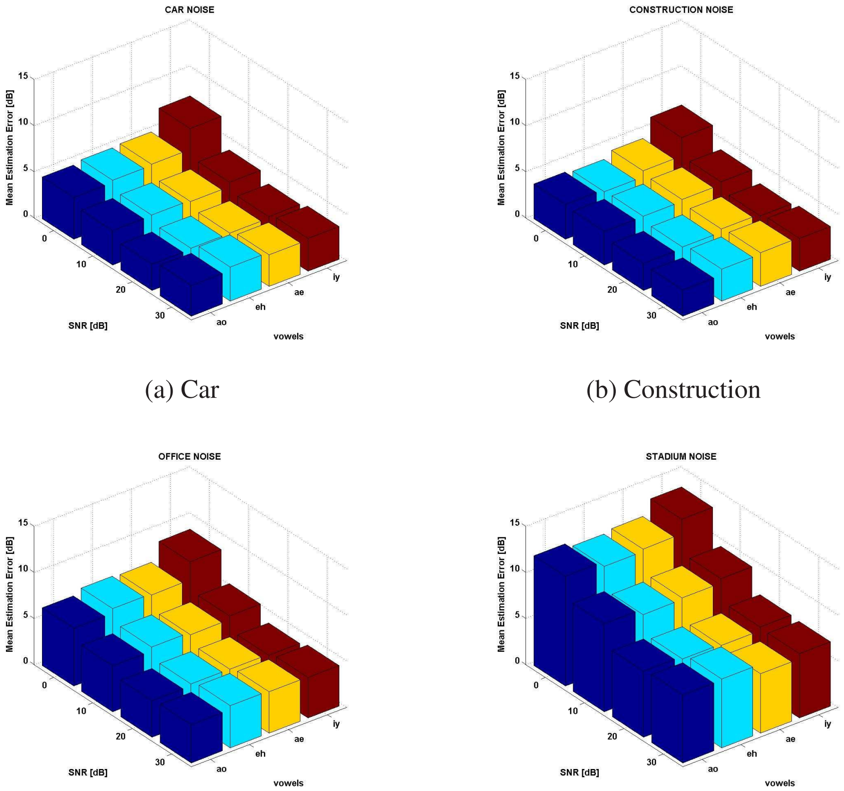
(d) Stadium

Figure 4.1: Average SNR estimation errors for the manual method

an average error between 3 and 5 dB of difference with the reference SNR.

A comparison of the errors reveals that the usage of the automatic method increases the average error by 1 dB in case of the Car, Construction and Office noises, while the increase is larger (between 2 and 5 dB) for the Stadium noise.

Even though the VAD impact on the SNR estimation depends on the type of noise, it however does not lead to heavily poorer performance because on



(a) Car

(b) Construction

(c) Office

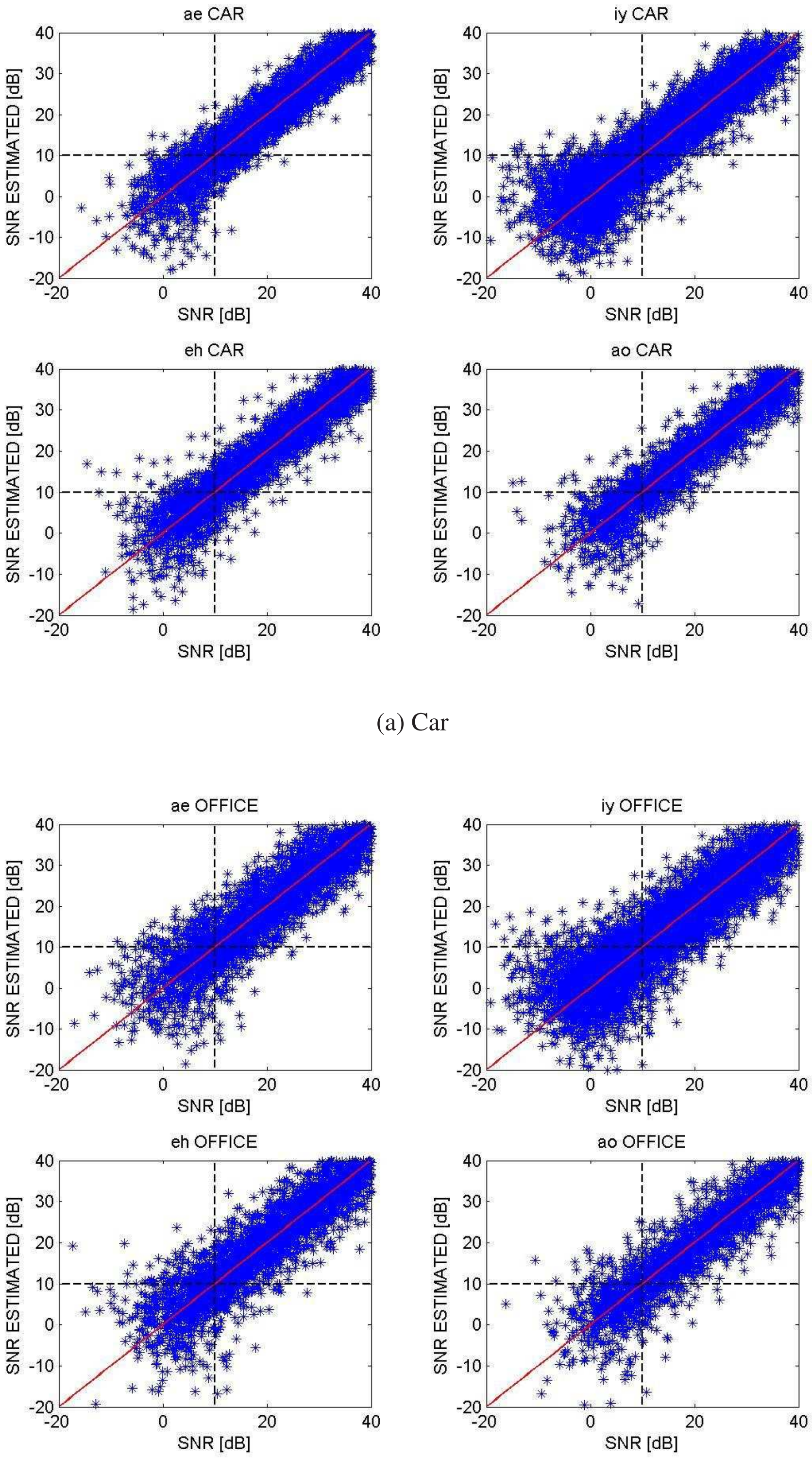
(d) Stadium

Figure 4.2: Average SNR estimation errors for the semi-automatic method

average the error grows by only 1-2 dB.

In both cases, when the reference SNR is 0 dB it can be seen that the "iy" vowel is subject to a high sensitivity for each kind of noise. The average estimation error generally is larger by 20-30% with respect to the other vowels.

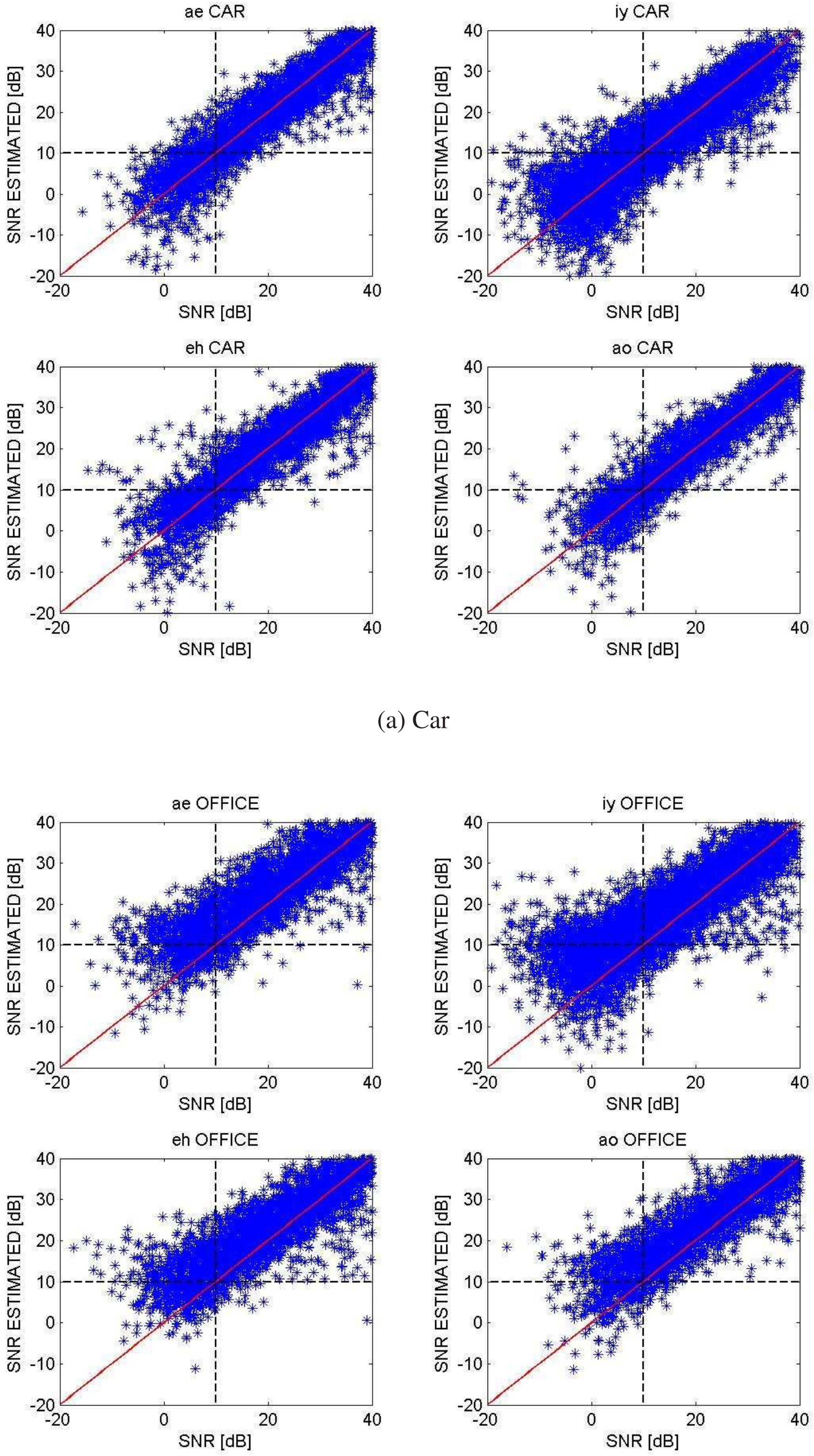
The plots in Figure 4.3 and Figure 4.4 show the correlation between the



(a) Car

(b) Office

Figure 4.3: Real vs. estimated SNR, manual method



(a) Car

(b) Office

Figure 4.4: Real vs. estimated SNR, semi-automatic method

real SNR and the estimated SNR for each of the 4 vowels in case of Car and Office noise. If we assume a critical threshold for rejecting a biometric sample of 10 dB, it is possible to outline 4 regions in each of these plots: the upper-left one, that encompasses data erroneously used because the SNR was over-estimated; the lower-right region, that comprises data erroneously discarded because the SNR was under-estimated, and the remaining regions (upper-right and lower-left), that contain data that were correctly discarded or used for the subsequent identity verification phases.

Tables 4.2 and 4.3, respectively, for manual and semi-automatic methods, show the error percentages depicted in Figure 4.3 and Figure 4.4. The semi-automatic method induces an increment of the percentage of low-quality data that is used for subsequent elaboration for the Office noise, while the percentages for the Car noise are similar to the ones of the manual method.

In the end, comparing the percentage of low-quality data erroneously used, it can be deduced that each vowel reacts in different ways: for instance, the “iy” vowel is one of the most robust. A similar comparison can be carried out in terms of high-quality data erroneously discarded.

4.4 Performance evaluation of Alize-LIA_RAL

In this section we present a study on how a speaker recognition system based on the Alize/LIA_RAL toolkit behaves when the data is affected by background noise. In particular, the section shows both the performance using a clean database and the robustness to the degradation of various natural noises, and their impact on the system. Finally, the impact of the duration to both training and test sequences is studied.

Car noise	**ae**	**iy**	**eh**	**ao**
Bad data used	15.39%	11.63%	15.34%	16.82%
Good data discarded	5.49%	6.75%	4.49%	4.91%
Office noise	**ae**	**iy**	**eh**	**ao**
Bad data used	22.14%	14.95%	21.76%	17.70%
Good data discarded	5.97%	7.97%	6.41%	6.00%

Table 4.2: Percentage errors for the manual method

Car noise	**ae**	**iy**	**eh**	**ao**
Bad data used	18.56%	15.42%	18.11%	18.77%
Good data discarded	4.94%	6.86%	4.61%	4.33%
Office noise	**ae**	**iy**	**eh**	**ao**
Bad data used	60.45%	42.70%	58.55%	56.46%
Good data discarded	2.35%	3.28%	1.53%	1.59%

Table 4.3: Percentage errors for the semi-automatic method

4.4.1 Speech and background noise database

For this experiment, we used the training portion of the DR1 TIMIT subset, that contains 38 people.

We generated the clean and noisy databases using the same protocol describe in Section 4.3.1.

4.4.2 Performance evaluation and results

In order to verify the performance of our system, we computed the genuine match scores and the impostor match scores for different types of noises and signal-to-noise ratio (SNR). The Detection Error Trade-off of each test case is shown in the following figures.

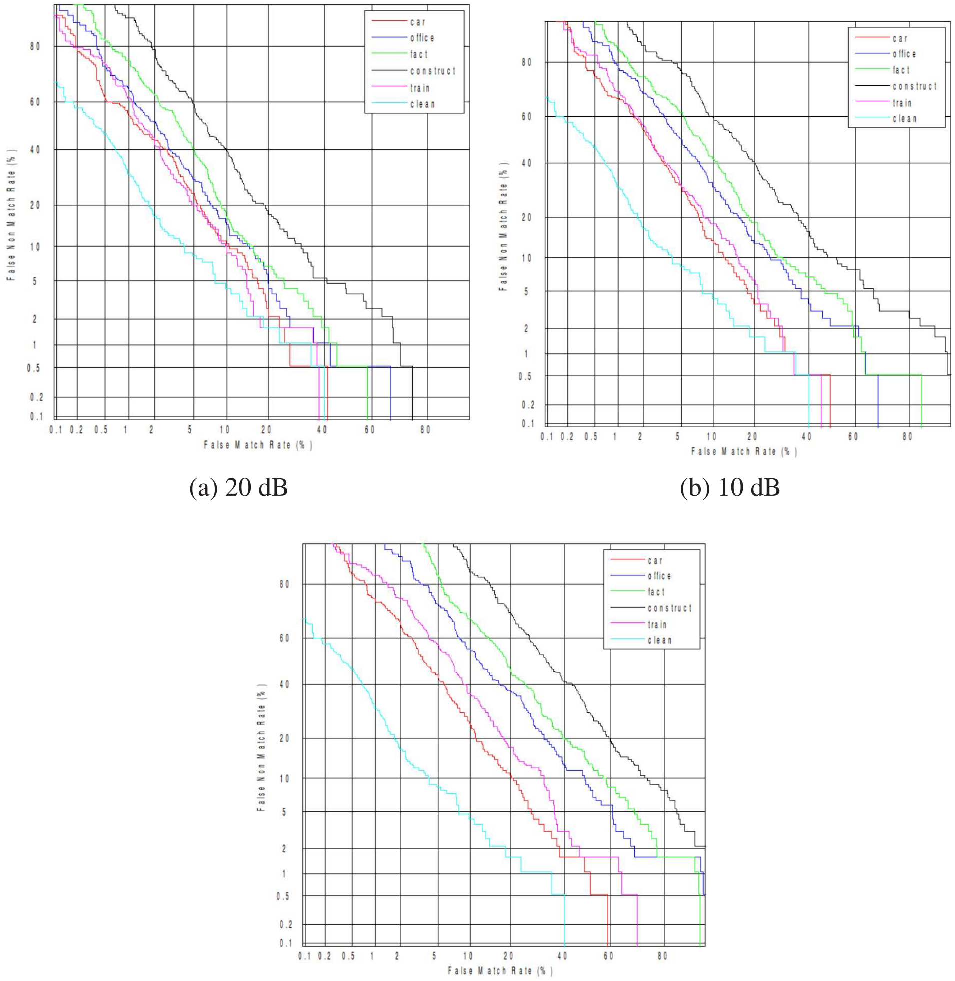
Figures 4.5 compare the performance on the basis of noise type for: (a) SNR=20 dB, (b) SNR=10 dB, (c) SNR=0 dB. In all cases we can notice major performance degradation after raising the noise level volume and a different impact on the system performance made by various noise types. In particular, car noise has less impact (EER=13 %) while construction noise is the most degrading noise type (EER=24 %). Algorithm performance in clean sequences points out an EER value of about 8 %, so the impact of the noise compromises the performance for EER percentage basis ranging from 5 to 15 %.

Another important result is the discovered relation about the recognition performance and the duration of the training and testing sequences. Figure 4.6a compares the DET achieved using clean sequences spanning the following durations:

- 2 training sequences (duration 6,24 sec), 2 true test sequences (duration 6,24 sec) and 2 false test sequences (6,24 sec);
- 8 training sequences (duration 25 sec), 2 true test sequences (6,24 sec) and 2 false test sequences (6,24 sec);

In this case the real impact of the training duration on the total system performance is evident.

Figure 4.6b shows the opposite case where a different duration of the test sequences is applied, in particular:



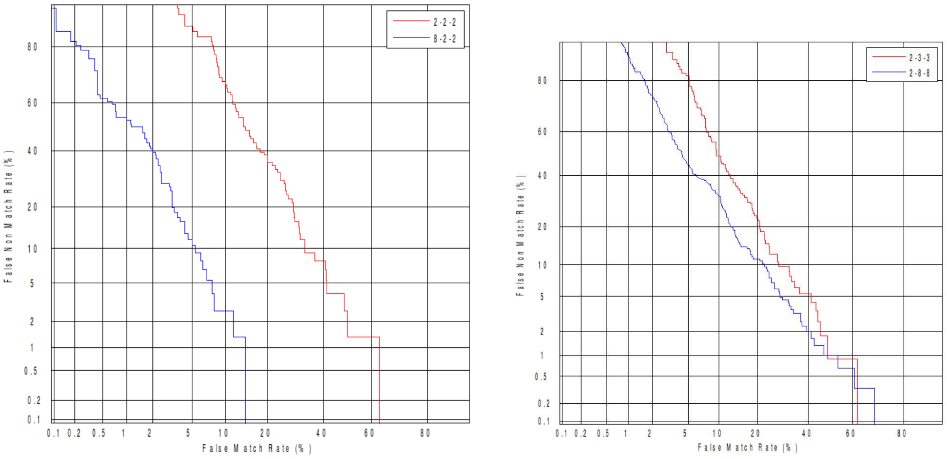
(a) 20 dB

(b) 10 dB

(c) 0 dB

Figure 4.5: DET vs. Noise type

- 2 training sequences (duration 6,24 sec), 3 true test sequences (duration 9,36 sec) and 3 false test sequences (9,36sec);



(a) 2-2-2 vs. 8-2-2

(b) 2-3-3 vs. 2-8-8

Figure 4.6: Training length vs. test length

- 2 training sequences (6,24 sec), 8 true test sequences (25 sec) and 8 false test sequences (25 sec).

In this case the different durations of the test sequences does not have much impact and the performance are very similar. Therefore, from this result it emerges that, for automatic speaker recognition, it is better to use longer duration sequences for training and shorter duration sequences for testing.

Finally, Figure 4.7 compares system performance in three different modalities: comparison of clean type training and testing sequences, comparison of clean training sequence and degraded testing sequence by car noise with SNR 0dB, and comparison of training and testing sequences both degraded by car noise with SNR 0dB. Analysing the three DET curves it is possible to see that employing one noisy sequence in the training phase does not contribute to the improvement of the performance, which remains similar to the clean-noisy case. Generally, we can therefore conclude that speaker identification

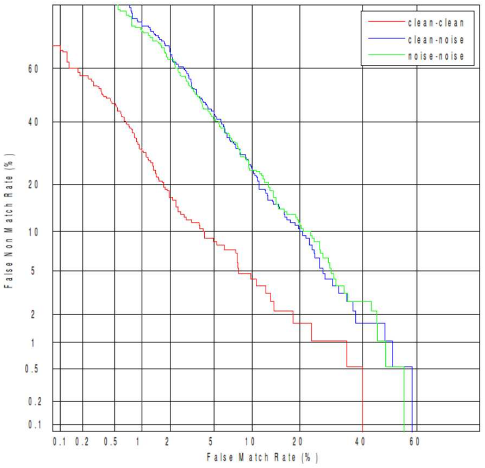


Figure 4.7: Clean-clean, Clean-Noisy, Noisy-Noisy

performance is sensitive to the degradation of one of the compared sequences (phonic test and testing).

4.5 The impact of voice activity detection

The performance of biometric speaker verification systems is largely dependent on the quality level of the input signal [41]. One of the most important components of such a system is the Voice Activity Detection (VAD) algorithm, as it has the duty of separating speech frames and noise frames, discarding the latter and feeding the speech frames to the rest of the system. This task becomes quite challenging as the Signal-to-Noise Ratio (SNR) of the input signal goes down [37] [38].

A VAD algorithm can use many techniques to classify speech and noise, such as an energy threshold or the analysis of the spectral characteristics of the audio signal. Due to these differences, different algorithms can behave differently in a given noise condition, and this is the reason for the study presented by this section.

The context in which we operate is the analysis of phone tappings in forensic investigations, and our task is to determine whether the conversation was carried on by a suspect or not. Those tappings are often noisy, so we generated from a speech database some audio files with the typical ON-OFF statistics of phone conversations and artificially added to them background noise in order to evaluate the performance of VAD algorithms and speaker identification at different SNR levels [36].

Our objective is to demonstrate that the usage of a single VAD is not the optimal solution, however, biometric identification performance can be improved by introducing a noise estimation component that can dynamically choose the best VAD algorithm for the estimated noise condition.

4.5.1 Alize/LIA_RAL

The speaker verification system used in this experiment is based on ALIZE/SpkDet , that is described in Appendix B.

Thanks to its modularity, we were able to run tests with the integrated VAD, described in Section 4.5.4, and with the other VAD algorithms (ideal and AMR), by converting the output of those algorithms in a format accepted by ALIZE/SpkDet .

4.5.2 The speech database

For our task we selected a TIMIT subset composed by 253 speakers, namely the union of DR sets 1, 2 and 3. Of those speakers, 63 were destined to train the UBM and 190 were used to train the identity models and to compute the match scores. With those speakers, we obtained 190 genuine match scores and 35910 ($190 \cdot 189$) impostor match scores for each simulation.

The speech files were used to generate two one-way conversation audio files, each containing speech material from 5 speech files and with an activity factor of 0.4, using the algorithm described in Section 4.5.3. In the case of the UBM speakers, both sequences were processed for the training phase, while in the case of identity models one sequence was used for the model training and the other was used for the computation of match scores.

The whole database was downsampled to 8kHz, to better match the forensic scenario, and normalized to an average power level of -26dB_{ovl}.

4.5.3 Generation of one-way conversations

In order to simulate the forensic scenario, and to give realistic input data to the VAD algorithms, we generated for each speaker two audio files that mimic one side of a two-people conversation, inserting speech and pauses according to the model described in ITU-T Recommendation P.59 “Artificial Conversational Speech” [42], that will now be briefly described.

According to this model, a conversation can be modelled as a Markov chain, whose state can be one of the following: A is talking, B is talking, Mutual silence, Double talk. A and B are the two simulated speakers.

The chain is depicted in Figure 4.8, along with the transition probabilities between the states.

The permanence in each of these states is given by the following equa-

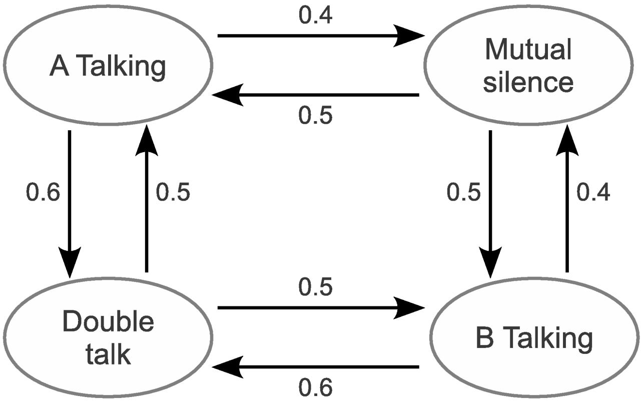


Figure 4.8: Markov chain used to generate the conversations

tions:

$$T_{st} = 0.854 \ln(1 - x_1)$$
$$T_{dt} = 0.226 \ln(1 - x_2)$$
$$T_{ms} = 0.456 \ln(1 - x_3)$$

where $0 < x_1, x_2, x_3 < 1$ are random variables with uniform distribution. T_{st} is the permanence time in the states in which a single speaker is talking, T_{dt} is associated to the double talk state and T_{ms} is used for the mutual silence state.

This model represents a two-way conversation, but we are interested in generating speech for one of the two sides of the conversation. So when the model is in the state “A is speaking” or “Mutual talk”, the generator adds speech material to the output sequence, while in the other two states the generator adds silence.

For this experiment, we used the Car, Office and Factory noises.

4.5.4 The LIA_RAL VAD

The LIA_RAL VAD is an energy-based off-line algorithm that works by training a GMM on the energy component of the input features. It then finds the Gaussian distribution with the highest weight w_i and uses its parameters to compute an energy threshold according to the following formula:

$$\tau = \mu_i - \alpha\sigma_i$$

where α is a user-defined parameter, and μ_i and σ_i are the parameters of the selected gaussian mixture Λ_i.

The energy threshold is then used to discard the frames with lower energy, keeping only the ones with a higher energy value.

4.5.5 The AMR VAD

The Adaptive Multi-Rate (AMR) Option 1 VAD [40] is a feature-based online algorithm that works by computing the SNR ratio in nine frequency bands, and decides which frames must be kept by comparing the SNRs to band-specific thresholds.

Note that this VAD is not optimized for speaker verification tasks, as it has the objective of minimizing the decision time, and it is designed to be used in real-time speech coding applications, while in a forensic biometric system the delay is not a significant parameter to minimize, and thus the VAD could use information from all the input signal to make its decision, as the LIA_RAL VAD does.

4.5.6 Evaluating VAD performance

In order to evaluate the VAD performance, we need to compare the results of the classification on a given input signal with a reference ideal classification that we know for sure to be correct.

In our experimental set-up, this ideal classification is derived by labelling the start and the end of speech segments generated by the model described in Section 4.5.3. This classification does not take into account pauses that can occur during the TIMIT spoken sentences, but it is a good approximation of an ideal classification.

The VAD classifier can misinterpret a given input frame in two ways: detecting a noise frame as speech (Noise Detected as Speech, *NDS*) or classifying a speech frame as noise (Speech Detected as Noise, *SDN*).

Those two errors are then further classified according to the position of the error with respect to the nearest word; see [43] for a discussion of those parameters.

For our analysis, we are not interested in the time when the misclassification occurs, as it is mainly useful when evaluating the perception effects of VAD errors [44], so we use the two NDS and SDN parameters, defined as follows for a single conversation:

$$\begin{aligned} NDS_{\%} &= \frac{N_{NDS} \cdot f}{C} \\ SDN_{\%} &= \frac{N_{SDN} \cdot f}{C} \end{aligned}$$

where N_{NDS} and N_{SDN} are, respectively, the number of NDS and SDN frames, f is the frame length expressed in seconds and C is the duration of the conversation expressed in seconds.

We then define a Total Error Rate (TER), as:

$$TER_{\%} = NDS_{\%} + SDN_{\%} \tag{4.2}$$

The TER is the percentage of audio frames that are misclassified by the VAD.

4.5.7 Experimental results

The starting point of our experiments is the creation of 9 noisy speech databases, obtained by summing to the one-way conversation speech database described in Section 4.5.3 the Car, Office and Factory noises, artificially setting the SNR to 20 dB, 10 dB and 0 dB.

Next the Equal Error Rate (EER) was computed over each database, first with the ideal segmentation and then by swapping this segmentation with the ones generated by the LIA_RAL VAD and by the AMR VAD, for a total of 27 simulations.

Finally, the VAD errors were computed using the metrics defined in Section 4.5.6.

In the clean case (Table 4.4), we reported the average $NDS_{\%}$, $SDN_{\%}$ and $TER_{\%}$, computed over all the speech samples used to run the speech verification simulations, one time for each VAD algorithm (including the ideal VAD).

In the noisy cases (Tables 4.5, 4.6, 4.7), since for each VAD the simulation was run once for each SNR level, the reported VAD error metrics are the average of the average of the value of each metric (denoted with μ) computed over all the speech samples, and their standard deviations σ are reported in order to better understand the nature of the data presented. Obviously, the VAD errors of the ideal VAD are always zero, so the standard deviation is omitted from the tables.

4.5.8 Analysis of the results

VAD algorithm	EER (%)	$\overline{NDS_{\%}}$	$\overline{SDN_{\%}}$	$\overline{TER_{\%}}$
ideal	**3.76**	0	0	0
AMR	4.36	1.08	4.81	5.89
LIA_RAL	3.77	4.33	31.74	36.07

Table 4.4: Results for clean speech

	EER (%)			**VAD Errors ($\mu \pm \sigma$, %)**		
VAD	0dB	10dB	20dB	$\overline{NDS_{\%}}$	$\overline{SDN_{\%}}$	$\overline{TER_{\%}}$
ideal	5.77	3.60	3.55	0	0	0
AMR	**4.95**	**4.77**	**3.38**	7.30 ± 1.55	4.88 ± 0.34	12.18 ± 1.89
LIA_RAL	6.49	5.43	4.87	3.95 ± 0.12	34.38 ± 0.24	38.33 ± 0.36

Table 4.5: Results table for CAR noise

Looking at Table 4.4, the first question is why the ideal segmentation yields an EER that is very close to the one that the LIA_RAL VAD obtained, in spite of a greater $\overline{TER_{\%}}$.

This is because the ideal segmentation does not focus only on vocalized sounds, that are known to carry the information that is needed to determine the identity of the speaker, but rather is an indicator of when the generator described in Section 4.5.3 introduced speech in the audio sequence. This therefore includes some sounds, like fricatives, that should be left out when

	EER (%)			**VAD Errors ($\mu \pm \sigma$, %)**		
VAD	0dB	10dB	20dB	$\overline{NDS_{\%}}$	$\overline{SDN_{\%}}$	$\overline{TER_{\%}}$
ideal	8.52	5.69	4.10	0	0	0
AMR	9.77	5.39	**3.75**	41.23 ± 5.24	5.08 ± 0.30	46.31 ± 5.53
LIA_RAL	**6.97**	**5.21**	4.00	0.83 ± 0.85	19.39 ± 3.52	20.22 ± 4.37

Table 4.6: Results table for OFFICE noise

	EER (%)			**VAD Errors ($\mu \pm \sigma$, %)**		
VAD	0dB	10dB	20dB	$\overline{NDS_{\%}}$	$\overline{SDN_{\%}}$	$\overline{TER_{\%}}$
ideal	7.84	5.01	4.70	0	0	0
AMR	7.27	**5.02**	**3.42**	13.64 ± 1.04	6.12 ± 1.33	19.76 ± 2.49
LIA_RAL	**6.58**	5.93	4.37	3.13 ± 1.44	16.87 ± 3.40	20.00 ± 4.84

Table 4.7: Results table for FACTORY noise

doing biometric identity comparisons. This also explains the worse performance of the ideal VAD in other cases like OFFICE, FACTORY 0 dB, etc. Analyzing the average errors made by the VAD algorithms, it is clear that the AMR VAD usually tends to be more conservative in the decision of rejection of speech, because its $\overline{NDS_{\%}}$ is always greater than LIA_RAL's; on the other hand, LIA_RAL always has a greater $\overline{SDN_{\%}}$ than AMR, and this means that it tends to be more selective in the decision of classifying a frame as noise.

The results for the CAR noise show that the AMR VAD always performs better than the LIA_RAL VAD in terms of EER, and it is supported by a significantly lower TER.

The OFFICE noise results do not show a clear winner between the two algorithms, as for high SNR the AMR VAD performs better, but as the SNR decreases, the LIA_RAL algorithm outperforms the AMR VAD. A similar pattern can be seen in the FACTORY results.

4.6 Conclusions and future work

In this chapter, we analyzed many of the problems that currently affect forensic speaker recognition. It is clear from the results of the previous sections that there is still no universal approach for speaker recognition in forensic context, and also that this applies to some of the smaller sub-problems.

More specifically, some ideas for future work in the SNR estimation area are:

- develop more effective SNR estimation algorithms, that can guarantee a lower average error and, most importantly, lower variance;
- estimate SNR in the sub-bands of interest of the main biometric indices adopted [36], typically in the fundamental frequency and in the first three formants;
- identify the critical SNR thresholds for each vowel and for each kind of noise by evaluating the impact of the noise on the whole identification process;
- use automatic environmental noise classifiers that allow to choose an SNR estimation model and critical thresholds tailored to the kind of noise [38] [45]

Regarding the selection of VAD algorithms, in the forensic context, where accuracy is truly important and results can be collected off-line, multiple VAD algorithms with different characteristics could be used, and all the identification decisions computed using them could then be fused using a majority rule or other fusion rules. In those critical kinds of analysis, it would be important that most of the decisions agreed between them, or else the disagreement could be an indicator that the choice of the VAD algorithm has a greater weight than desired.

More broadly, based on the results of the research work described in this chapter, it is clear that both the SNR estimation and VAD algorithm selection problems could benefit from an adaptive approach that first estimates the characteristics of background noise and then select the algorithm that performs better in that context [46].

CHAPTER

FIVE

MULTI-MODAL WEB AUTHENTICATION

5.1 Introduction

In the last 20 years, thanks to the development of the World-Wide Web (WWW), many of the activities carried on the Internet have been increasingly moved to web browsers and web sites, to the point that most people today commonly misidentify the Internet for the WWW itself [47].

The Web browser is increasingly becoming the access point for most online activities, and the recent trend of moving from traditional client-server computing paradigm to cloud-based computing, coupled with the rise of thin client for the WWW (netbooks), will only contribute towards this change of perspective, making web browsers more and more the entry point of the Internet for most people.

In this context, biometric authentication systems will need to adapt themselves to be successful. In particular, they should not depend on external devices that are not commonly found in consumer hardware, and they should be accessible from inside a Web browser. Moreover, given that the quality of the samples acquired from consumer hardware - think about the microphone or the webcam of a netbook - is not comparable to the quality of samples acquired with professional biometric devices, it is possible that those systems will need to exploit more than one biometric trait to be effective.

In this chapter, we will describe the architecture and the implementation of Biometric4Net[1] [48], a prototype of web-based multi-modal biometric authentication system. The system is open source, and the performance of its biometric back-ends were evaluated using a non-chimeric multi-modal biometric database called UCT10.

This chapter is structured as follows: in Section 5.2 we will describe the Biometric4Net architecture and implementation choices; in Section 5.3 we will describe the UCT10 multi-biometric database; in Section 5.4 we will describe the performance evaluation of the biometric backends; finally, in Section 5.5 we will draw our conclusions about the system.

5.2 Proposed architecture

The prototype system is composed by two sub-systems, the biometric authentication server (from now on simply referred to as the *server*) and the web-based user interface, that will be simply called *client*.

Figure 5.1 shows the proposed architecture, along with some hints on the appearance of the client. Being it a web-based architecture, the two clients will have to communicate using the HTTP protocol. The client will need to run inside a web browser, and the server will either need an external web server or implement it by itself.

[1] this work was supported by the GARR Consortium with Grant 2009-01

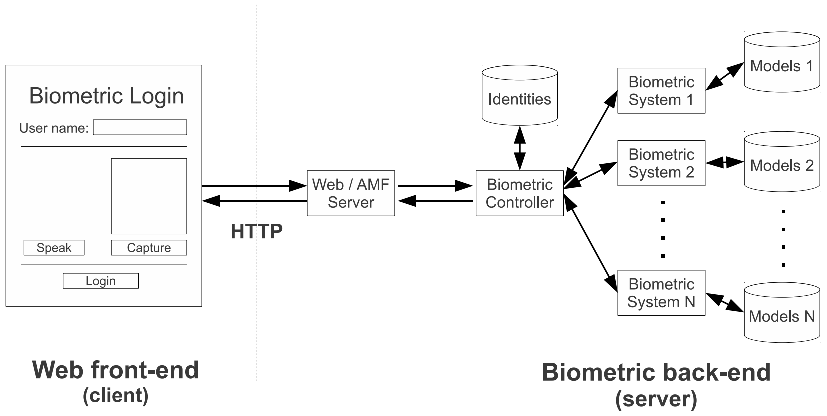


Figure 5.1: The proposed architecture

In the remainder of this section, we will describe the structure of both the components of the system, and discuss briefly the alternatives considered while choosing the technologies used for the prototype implementation.

5.2.1 The client

The objective of the web client is to let the user interact with the whole system, allowing him to:

- declare his identity;
- give his biometric data to the system;
- receive feedback from the system;

- if needed, have access to the requested resource.

The identity declaration, feedback communication and resource access tasks are trivial to implement, and very common in web applications. The only task that is worth of discussion is the acquisition and encoding of biometric data, that implies access to the relevant devices of the user's computer, in particular the microphone and the webcam.

One of the most common ways of achieving this task is the development of browser plug-ins that bypass the protection of the browser and have direct access to the operating system. An example of this kind of solutions is the family of ActiveX browser plug-ins, for the Microsoft Internet Explorer browser. This kind of solution has the huge disadvantage of being tied to a single browser (and operating system), greatly limiting the portability of the system.

Given that one of the objectives of the system is to be open, and not only in terms of giving access to the source code, this family of solutions was not considered.

After a careful analysis, three technologies were deemed worthy of consideration for this task:

1. HTML 5;

2. Java;

3. Flash;

In the early stages of investigation, the still-evolving HTML 5 specifications [49], the `<device>` item was said to allow client-side web applications to access the user's devices, including audio/video and USB devices.

This would have been the ideal solution: standards-compliant, open, cross-browser. Unfortunately, the specification is still not complete and not fully implemented by all the leading browsers, thus the adoption of HTML 5 would paradoxically result in a limited usability of the application. The choice of not using HTML 5 was later revealed to be appropriate, because the `<device>` tag was removed from the specification, in favour of the `getUserMedia` API [50].

Both Java and Flash allow web applications to have access to the microphone and to the webcam of the user, and they both offer to developers a free SDK. It is widely accepted that Java does not offer the same cross-platform compatibility level of Flash, and moreover Flash it is already in use in many audio/video communication platforms.

This is why the technology chosen for the development of the client platform is Flash, using as a development framework the Flex 4 SDK by Adobe [51].

5.2.2 The server

Having chosen the front-end technology, we can now focus on the communication technique between the front-end and the back-end. There are two possible strategies:

- send data in real-time;
- record data in the client and send it when the recording phase is finished.

For biometric authentication purposes, the two strategies are indifferent, as the biometric score has to be computed on the whole data segment acquired from the user: all the data must be processed before the final decision. So the choice can be driven only by implementation-related reasons.

The streaming solution can be implemented using the Real-Time Messaging Protocol (RTMP), a protocol that is well-integrated into the Flash platform. Unfortunately, the most common RTMP server is Adobe Flash Media Server (FMS), a closed-source and commercial software. Luckily, there are some open source alternatives, that have all been tested by us:

- Red5 [52];
- erlyvideo [53];
- rtmplite [54];

The first, Red5, is the most complete and most complex of the three servers, and it is the one with which most of our exploratory tests have been conducted. Once Red5 was set up, the other two servers were briefly installed and configured. While once the configuration phase has been done the usage of the protocol is relatively easy, it was clear that an RTMP server is an additional layer of complexity that could be avoided, with the additional constraint that all the acquisition and pre-processing had to be done on the client, leaving to the server only the biometrics-related tasks.

The next step was then the investigation on how the data should be encoded and serialized. The most natural choice is the adoption of the AMF (Action Message Format) serialization format, used primarily by Actionscript (one of the main languages of the Flash platform) but with bindings for many other languages.

Finally, for the code of the server component we chose to adopt the Python programming language, due to its flexibility, to the wide availability of lots of third-party libraries, and to its dynamic typing that makes it perfectly suited for a relatively small portion of code that acts as middleware between many systems.

The server component comprises two layers:

- A communication layer, based on PyAMF and WSGI;

- A middle layer that encodes data for each biometric backend;

In this prototype, only the voice backend has been implemented, while the performance evaluation tests were conducted on both voice- and face-based biometric recognition systems, as the whole encoding and transmission process introduced by the client/server system does not have any effect on the biometric performance of the backends.

5.2.3 The authentication process

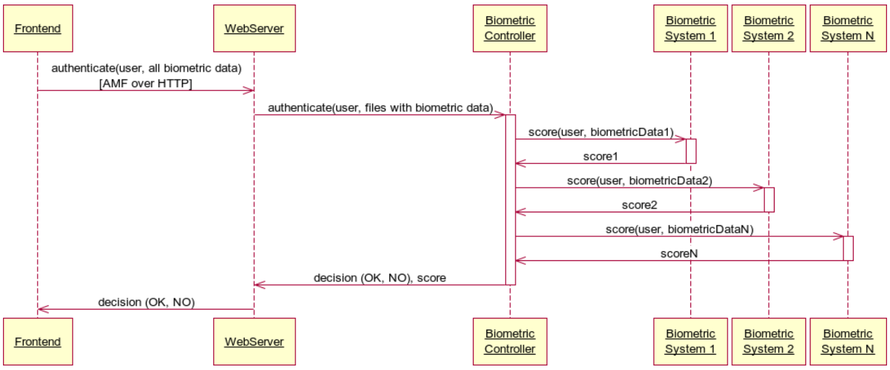


Figure 5.2: Sequence diagram of the authentication process

Figure 5.2 shows a sequence diagram of the authentication process, that we will comment with the help of the screenshots of the prototype shown in Figure 5.3.

When the user wants to gain access to a (fictitious) resource protected by Biometric4Net, as a first step he has to declare his identity (Figure 5.3a). This phase is implemented as a simple selection from a combo-box containing all the identities presented in the database, but it could very easily be replaced by the insertion of an user name. Also, currently the only identity validation factor is the biometric data, but it would be easy to add a second authentication factor, such as a password.

Next, the user has to start the biometric acquisition. The security model of Flash imposes that he has to explicitly authorize the Flash application to gain access to the webcam and the microphone the first time that he uses it, and this is shown in Figure 5.3b. After the authorization, the user records his sentence and then stops the acquisition, as shown in Figure 5.3c.

At this point, the process summarized in Figure 5.2 starts. The client sends to the server the identity verification request, serializing to AMF the tuple (claimed identity, biometric data). The web server receives the HTTP request and processes it, calling via RPC the `authenticate` method of the biometric Controller.

The Controller is the component that queries the biometric systems and employs a score-level fusion algorithm to decide whether the claim must be considered true or not. Most fusion algorithm have already been implemented in the Controller, while only the backend for Alize/LIA_RAL has been implemented.

Each biometric system can be located in the same machine of the Controller or in another machine. This is totally transparent to both the User and the Controller.

As soon as the decision is taken by the Controller, it returns a tuple (decision, score) of types (boolean, float) to the Web Server, which simply proxies those value to the Client via an HTTP response. The Client has then the responsibility of rendering the choice to the User, as shown in Figure 5.3d, and let him access the protected resource if the decision is positive.

5.3 The UCT10 database

In order to evaluate the performance of multi-modal biometric systems, it is necessary to have a database that contains biometric data for each of the combined modalities. Ideally, to study the interaction between the different traits, for each person in the database there should be one or more biometric sample for each trait.

Unfortunately, this is not always the case, as there are practical difficulties in acquiring multi-modal databases. Usually, it is common practice among multi-modal biometric researchers to build artificial databases by arbitrarily associating biometric data acquired from different people into one *artificial person*. These databases are called *chimeric databases*, and recent studies have shown that the hypothesis of statistical independence between the traits does not hold [55].

This is why, in order to evaluate the performance of Biometric4Net, we built a true multi-modal biometric database, called UCT10 (University of Catania - 2010).

The database is composed of 50 people, and from each of them we acquired:

- 4 speech samples, of the average duration of 30 seconds;
- 10 frontal face pictures;
- 5 real signatures and 5 unskilled signature forgeries;
- 2 heart sound sequences of the average duration of 60 seconds.

Each person who contributed to the database has signed an agreement that authorizes the department to which we belong, the Dipartimento of Ingegneria

Elettrica, Elettronica and Informatica of the University of Catania, to use this data for research purposes and, if needed, to give this data to third parties, also for research purposes. In no way this data is associated with the actual identity or name of each of them.

5.4 Performance evaluation

In this section, we will describe the results of a performance evaluation of three biometric systems based on three different biometric traits and three score-level fusion algorithms, to justify the multi-modal approach and to show that in our context the more appropriate set of biometric traits appears to be the one composed by voice and face.

The comparison have been carried on using the UCT10 database, that has been described in Section 5.3.

The first biometric system is Alize/LIA_RAL, that is based on voice. It is described further in Appendix B.

The second and the third systems are 2DFace and Get-Int HMM. The first one is based on face recognition and the second on dynamic signature recognition. Both are described in [56].

We selected those three biometric traits (voice, face and signature dynamics) because they seem to be the ones that are more easily acquirable using widely available consumer-grade hardware that is accessible through software running in a web browser.

The first test, whose results are summarized in Table 5.1, have the objective to simply compare the recognition performance of the three systems on the UCT10 database.

It is easy to see that the Voice and Face systems give similar results

Biometric trait	**System**	**EER (%)**
Voice	ALIZE/SpkDet	4.25
Face	2DFace	**4.06**
Signature	Get-Int	12.65

Table 5.1: Results of the mono-modal identity verification tests

(4.25 % vs. 4.06 % EER), while the signature system has a considerably higher EER of 12.65 %. Note that this does not necessarily mean that multimodal systems using the signature will perform worse than systems using the other traits, as the biometric traits are not orthogonal and we can expect some correlation effects.

Biometric traits	**sum rule, EER (%)**	**product rule, EER (%)**
Voice/Face	0.43	**0.40**
Voice/Signature	4.12	4.12
Face/Signature	3.25	2.76
Voice/Face/Signature	**0.36**	0.78

Table 5.2: Results of the multi-modal identity verification tests

This is shown more clearly in Table 5.2, where the scores of the systems were fused in all their possible combinations using two different fusion techniques, the sum rule and the product rule.

The best value of EER is obtained by fusing the scores of all the three systems using the sum rule. This clearly shows that each of the systems gives its contribution to improving the accuracy of the global system.

In the choice of the traits to use in an hypothetic biometric system, the recognition performance is only one of the parameters of the choice. In this particular case, we would have suggested to use the Voice/Face fusion using the product rule. While this combination does not offer the absolute best performance, it is pretty close to the best one and the difference (0.04 % EER) is really negligible when doing performance evaluation on databases that are so small. Moreover, this set of biometric traits does not require a separate device for the acquisition of the signature and is probably perceived by the user as an acceptable couple of biometric traits.

Finally, a consideration on the type of fusion: in this analysis we considered only parallel fusion strategies; this means that all the data must be collected before the recognition process begins. Another class of fusion strategies is the one that employs serial fusion. This means that biometric samples are acquired from the user only if needed. If the system has a significant confidence on the analysis performed on the first biometric sample, the user will not be requested to provide other samples and the decision will be final even with only one sample.

In certain scenarios, these strategies can greatly improve the usability of the biometric samples, and should be considered when designing real-world biometric systems.

5.5 Conclusions and future work

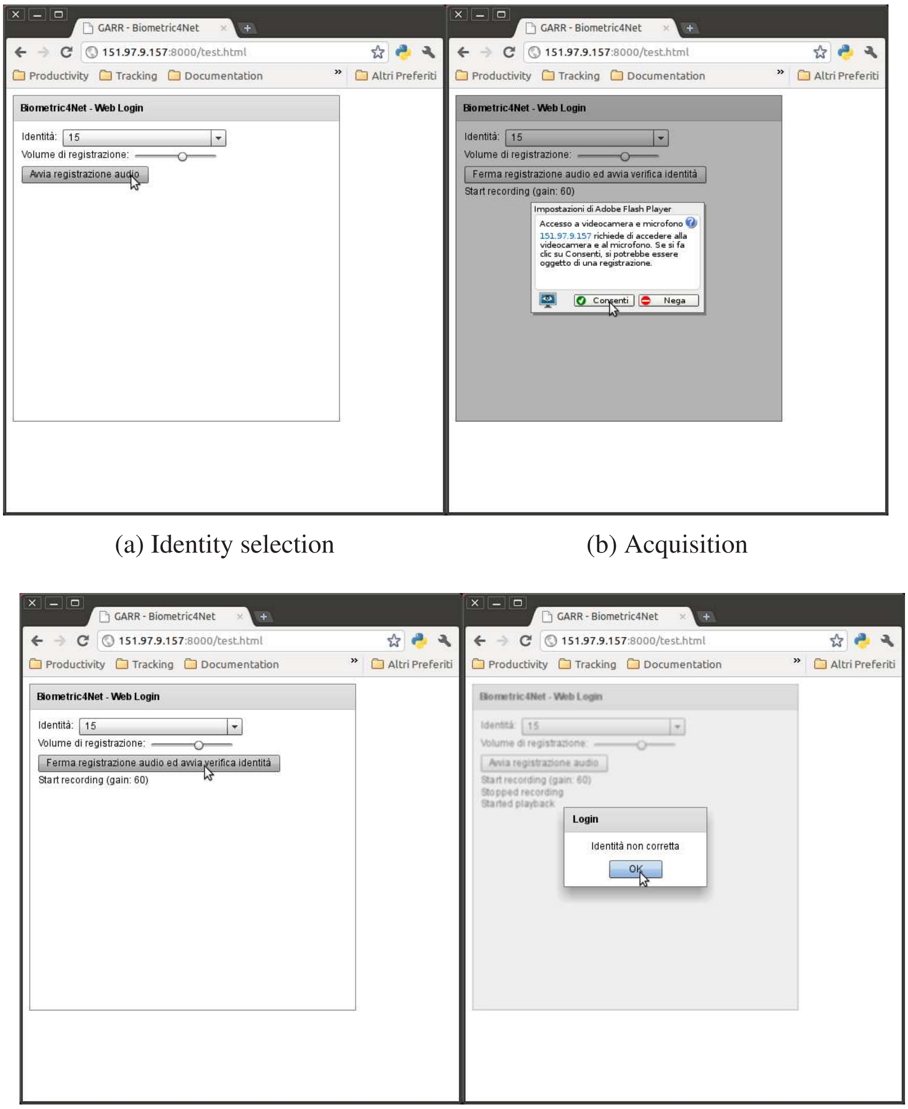
In this chapter, we presented Biometric4Net, an open source prototype of web-based multi-modal biometric authentication system.

We discussed its client/server architecture, the design and implementation choices that have driven its development and a performance evaluation of the biometric backends in mono- and multi-modal setups, over the multi-modal biometric database UCT10.

The system is freely available and can be used by the community as a starting point for the implementation of research multi-biometric systems. Some work would be required to turn it into a real-world application, for instance:

- encryption and cryptographic authentication of all the client/server communications;
- implementation of an enrolling interface;
- integration with directory services like LDAP and authentication services like Kerberos;

Nonetheless, the system can still be used in research setups, and it can also easily be extended to become a distributed biometric samples collection engine, to help the biometric research community in building larger multi-modal non-chimeric biometric databases.



GARR - Biometric4Net
151.97.9.157:8000/test.html
Productivity
Tracking
Documentation
Altri Preferiti
Avvia registrazione audio
Ferma registrazione audio ed avvia verifica identità
Consenti
Nega
OK

Biometric4Net - Web Login

Identità: 15

Volume di registrazione:

(a) Identity selection

Start recording (gain: 60)

Impostazioni di Adobe Flash Player

Accesso a videocamera e microfono

151.97.9.157 richiede di accedere alla videocamera e al microfono. Se si fa clic su Consenti, si potrebbe essere oggetto di una registrazione.

(b) Acquisition

(c) End of acquisition

Start recording (gain: 60)
Stopped recording
Started playback

Login

Identità non corretta

(d) Decision

Figure 5.3: Authentication process screenshots

CHAPTER

SIX

CONCLUSIONS

The recent trends in research activity and industrial product development show that biometric recognition techniques will gain more and more importance as time passes by. In this thesis, we have analyzed three different aspects of biometric systems.

Heart sounds are a new and promising biometric trait that has unique properties of universality and spoof robustness; in this thesis, we presented two systems based on this modality, the most performing on which is based on a statistical approach that uses Gaussian Mixture Models to represent the biometric templates; we also introduced some new algorithms for processing heart sounds and a new time-based feature set. The performance of this trait are still far from the conventional biometrics, but its unique properties, the increasing trends of better performance and the increasing number of research works suggest that, when it will reach a more mature stage, it might be helpful in high-security multi-biometric systems.

Automatic text-independent speaker recognition is a technique that would help avoiding subjective mistakes made during trials; in this thesis, we analyzed many of the problems that the forensic context adds to conventional speaker recognition, namely how noise affects the different components of the system, and we highlight some possible solutions for mitigating those problems.

The rise of web-based services offer a new use-case for remote authentication based on biometrics. The usage of consumer-grade hardware for the acquisition process leads to a degradation of the sample, thus making it necessary to use multi-biometric techniques. In this thesis, we discussed the architecture and implementation of an open source web-based multi-biometric system called Biometric4Net; three different biometric traits were considered for this system, and after analyzing the performance of all the 14 possible combination of these traits using two different score-level fusion rules, we demonstrated that the usage of speaker and face recognition, fused with the score product rule, yields the best compromise between good performance and ease of use.

APPENDIX

A

GAUSSIAN MIXTURE MODELS

In this chapter, we present a brief overview of the concepts and equations that are behind Gaussian Mixture Models, a statistical method for the approximation of complex data distributions.

In the field of biometric recognition, one of the most well-known fields of application of GMMs is speaker recognition [41], but in this thesis the same approach, with the due adaptations, has been applied also to heart-sounds biometry.

A GMM λ is a mixture composed by a weighted sum of N Gaussian probability distributions. The i-th Gaussian PDF is defined by the following formula:

$$p_i(x) = \frac{1}{(2\pi)^{D/2}\,|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)'\Sigma_i(x-\mu_i)}$$

where μ_i and Σ_i are respectively the mean vector and the covariance matrix of p_i.

The probability that a given D-dimensional vector x derives from the λ GMM is:

$$p(x|\lambda) = \sum_{i=1}^{N} w_i p_i(x) \tag{A.1}$$

Where w_i is the weight of each individual Gaussian. So the λ GMM model is defined by the following parameters:

$$\lambda = \{w_i, \mu_i, \Sigma_i\} \tag{A.2}$$

Those parameters are learned via the Expectation Maximization (EM) algorithm [57] in the model training phase, that in a biometric system is the enrollment phase.

The application of GMMs to speaker recognition has led to the development of the GMM/UBM method [41], where UBM stands for Universal Background Model.

The idea beyond this technique is to model separately the speaker and the set of people that, in our context are *not* the speaker (the world model).

In this discussion, we will not talk about *speakers* but of generic *identities*, and instead of *speech samples* we will talk about *signals*, generalizing the method.

Given an input signal s and a stated identity I, the problem of determining whether s belongs to I (represented by its model λ_I) is equivalent to a hypothesis test between two hypotheses:

$$\begin{aligned} &H_0 : s \text{ belongs to } I \\ &H_1 : s \text{ does not belong to } I \end{aligned}$$

This decision can be taken using a likelihood test:

$$S(s,I) = \frac{p(s|H_0)}{p(s|H_1)} \begin{cases} \geq \theta \text{ accept } H_0 \\ < \theta \text{ reject } H_0 \end{cases} \tag{A.3}$$

where θ is a decision threshold determined by the context in which the system is deployed.

We model the probability $p(s|H_0)$ using Gaussian Mixture Models.

The input signal is converted by the front-end algorithms to a set of K feature vectors, each of dimension D, so we can write:

$$p(s|H_0) = \prod_{j=1}^{K} p(x_j|\lambda_I) \tag{A.4}$$

In order to compute the score (A.3) that must be compared to the system's threshold, we still need to estimate $p(s|H_1)$.

In the GMM/UBM framework, this probability is modelled by building a speaker model trained with a set of speakers that represent the variability of the people that might use the system, the UBM.

The final score of the identity verification process, expressed in terms of log-likelihood ratio, is

$$\Lambda(s) = \log S(s,I) = \log p(s|\lambda_I) - \log p(s|\lambda_W) \tag{A.5}$$

APPENDIX

B

ALIZE/LIA_RAL

All the GMM-based experiments that have been described in this thesis were implemented using the free (as in free speech) speaker recognition toolkit Alize/LIA_RAL [58, 59].

This toolkit is developed jointly by the members of the ELISA consortium [60], and consists of two separate components: Alize, that is the low-level statistical framework, and LIA_RAL, that is the set of high-level utilities that perform each of the tasks of a state-of-the-art speaker recognition system. One of its main advantages is the high level of modularity of the tools: each program does not directly depend on the others and the data between the modules is exchanged via text files whose format is simple and intuitive. This means that researchers can easily change one of the components of the system with their own program, without having to modify its source code but only adhering to a set of simple file-based interfaces.

In this appendix, we will briefly describe all the components of a typical experiment that uses the ALIZE/SpkDet toolkit.

B.1 Feature extraction

LIA_RAL does not contain any module that performs feature extraction; all the experiments in this thesis used the Speech Signal Processing Toolkit (SPro) [26] for feature extraction tasks. SPro allows to extract different types of features, using filter-banks, cepstral analysis and linear prediction.

A notable exception is the computation of the FSR (see Section 3.3.2), that is done using custom programs written in GNU Octave or C++; one of them is built using the Alize framework, and has the duty of modifying the feature files containing cepstral parameters by adding the FSR feature to each vector; the other one computes the average FSR by analyzing the input signal.

B.2 Frames selection

The second step of the recognition process is to remove the frames that do not carry useful information. When dealing with the speech signal, this task is carried on by VAD algorithms, that have been already described in Section 4.5. Each of these VAD algorithms was implemented by a different program, and their output was always converted to a format that is compatible with the LIA_RAL toolkit.

For the heart sound, the S1/S2 segmentation algorithm described in Section 3.3.2 was used as the equivalent of VAD.

The default VAD algorithm in LIA_RAL, described in Section 4.5.4, is implemented in the utility *EnergyDetector*.

B.3 Feature normalization

The third step is the feature normalization, that changes the parameters vectors so that they fit a zero mean and unit variance distribution. The distribution is computed for each file.

The tool that performs this task is called *NormFeat*.

B.4 Models training

To use the UBM/GMM method, it is necessary first to create a world model (UBM), that represents all the possible alternatives in the space of the identities enrolled in the system; then, from the UBM, the individual identity templates are derived from the UBM using the Maximum A-Posteriori (MAP) estimation algorithm.

The tool used for the computation of the UBM is *TrainWorld*, while the individual training models are computed using *TrainTarget*.

B.5 Scoring

The computation of scores is done via the *ComputeTest* program, that scores each feature set against the claimed identity model and the UBM, and gives as output the log-likelihood ratio.

In order to take a decision, the system has then to compare the score with a threshold, and then accept or reject the identity claim. The decision step is implemented in LIA_RAL utility *Scoring*, but in this thesis we have not used it.

BIBLIOGRAPHY

[1] A. K. Jain, A. A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, pp. 4–20, Jan 2004.

[2] S. Prabhakar, S. Pankanti, and A. K. Jain, "Biometric recognition: Security & privacy concerns," *IEEE Security and Privacy Magazine*, vol. 1, no. 2, pp. 33–42, 2003.

[3] A. K. Jain, A. A. Ross, and S. Pankanti, "Biometrics: A tool for information security," *IEEE Transactions on Information Forensics and Security*, vol. 1, pp. 125–143, Jun 2006.

[4] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*. Springer, 2008.

[5] A. A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of Multibiometrics*. Springer, 2009.

[6] A. K. Jain and A. Ross, *Introduction to Biometrics*, ch. 1, pp. 1–22.

[7] M. Leotta, F. D. Natale, and A. Spadaccini, “Cashma: Rapporto tecnico sullo stato dell’arte delle tecniche biometriche mono-modali,” tech. rep., CC ICT-SUD, Catania, June 2011.

[8] A. Rubin, “Unwrapping ice cream sandwich on the galaxy nexus: http://googleblog.blogspot.com/2011/10/unwrapping-ice-cream-sandwich-on-galaxy.html,” 10 2011.

[9] T. Kikkunen and H. Li, “An overview of text-independent speaker recognition: from features to supervectors,” *Speech Communication*, 2009.

[10] J. Fierrez-Aguilar, N. Alonso-Hermira, G. Moreno-Marquez, and J. Ortega-Garcia *Proceedings of BIOAW*, 2004.

[11] R. Plamondon and S. N. Srihar *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[12] A. Ross, K. Nandakumar, and A. K. Jain, *Introduction to Multiiometrics*, ch. 14.

[13] L. Biel, O. Pettersson, L. Philipson, and P. Wide, “Ecg analysis: A new approach in human identification,” *IEEE Transactions on Instrumentation and Measurement*, vol. 50, pp. 808–812, Feb 2001.

[14] F. Beritelli and A. Spadaccini, *Human Identity Verification based on Heart Sounds: Recent Advances and Future Directions*, ch. 11, pp. 217–234. June 2011.

[15] M. Sabarimalai Manikandan and K. Soman, “Robust heart sound activity detection in noisy environments,” *Electronics Letters*, vol. 46, pp. 1100 –1102, 5 2010.

[16] F. Beritelli and S. Serrano, “Biometric Identification based on Frequency Analysis of Cardiac Sounds,” *IEEE Transactions on Information Forensics and Security*, vol. 2, pp. 596–604, Sept. 2007.

[17] L. Rabiner, R. Schafer, and C. Rader, “The chirp z-transform algorithm,” *Audio and Electroacoustics, IEEE Transactions on*, vol. 17, pp. 86 – 92, June 1969.

[18] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 357–366, Aug 1980.

[19] F. Beritelli and A. Spadaccini, “Human Identity Verification based on Mel Frequency Analysis of Digital Heart Sounds,” in *Proceedings of the 16th International Conference on Digital Signal Processing*, July 2009.

[20] K. Phua, J. Chen, T. H. Dat, and L. Shue, “Heart sound as a biometric,” *Pattern Recognition*, vol. 41, pp. 906–919, Mar 2008.

[21] D. H. Tran, Y. R. Leng, and H. Li, “Feature integration for heart sound biometrics,” in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 1714 –1717, 2010.

[22] J. Jasper and K. Othman, “Feature extraction for human identification based on envelogram signal analysis of cardiac sounds in time-frequency domain,” in *Electronics and Information Engineering (ICEIE), 2010 International Conference On*, vol. 2, pp. V2–228 –V2–233, 2010.

[23] S. Fatemian, F. Agrafioti, and D. Hatzinakos, “Heartid: Cardiac biometric recognition,” in *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, pp. 1 –5, 2010.

[24] N. El-Bendary, H. Al-Qaheri, H. M. Zawbaa, M. Hamed, A. E. Hassanien, Q. Zhao, and A. Abraham, “Hsas: Heart sound authentication system,” in *Nature and Biologically Inspired Computing (NaBIC), 2010 Second World Congress on*, pp. 351 –356, 2010.

[25] F. Beritelli and A. Spadaccini, “Heart sounds quality analysis for automatic cardiac biometry applications,” in *Proceedings of the 1st IEEE International Workshop on Information Forensics and Security*, Dec. 2009.

[26] G. Gravier, “SPro: speech signal processing toolkit,” 2003.

[27] F. Beritelli and A. Spadaccini, “A statistical approach to biometric identity verification based on heart sounds,” in *Proceedings of the Fourth International Conference on Emerging Security Information, Systems and Technologies (SECURWARE2010)*, pp. 93–96, IEEE, Jul 2010.

[28] F. Beritelli and A. Spadaccini, “An improved biometric identification system based on heart sounds and gaussian mixture models,” in *Proceedings of the 2010 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications*, pp. 31–35, IEEE, Sep 2010.

[29] A. Spadaccini and F. Beritelli, “Performance Evaluation of Heart Sounds Biometric Systems on an Open Dataset,” in *to appear in Proceedings of the 5th IAPR International Conference on Biometrics*, July 2012.

[30] F. Beritelli, S. Casale, R. Grasso, and A. Spadaccini, “Performance evaluation of snr estimation methods in forensic speaker recognition,” in *Proceedings of the Fourth International Conference on Emerging Security Information, Systems and Technologies (SECURWARE2010)*, pp. 88–92, IEEE, Jul 2010.

[31] F. Beritelli and A. Spadaccini, “Performance evaluation of an automatic forensic speaker recognition system based on gmm,” in *Proceedings of the 2010 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications*, pp. 22–25, IEEE, Sep 2010.

[32] F. Beritelli and A. Spadaccini, “The role of voice activity detection in forensic speaker verification,” in *Proceedings of the 17th International Conference on Digital Signal Processing (DSP2011)*, pp. 1–6, IEEE, Jul 2011.

[33] P. Rose, *Forensic Speaker Recognition*. Taylor and Francis, 2002.

[34] J. Richiardi and A. Drygajlo, “Evaluation of speech quality measures for the purpose of speaker verification,” in *Proceedings of Odyssey, The Speaker and Language Recognition Workshop*, 2008.

[35] M. Falcone, A. Paoloni, and N. D. Sario, “Idem: A software tool to study vowel formant in speaker identification,” in *Proceedings of the ICPhS*, pp. 145–150, 1995.

[36] F. Beritelli, “Effect of background noise on the snr estimation of biometric parameters in forensic speaker recognition,” in *Proceeding of the International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2008.

[37] F. Beritelli, S. Casale, and S. Serrano, “A low-complexity speech-pause detection algorithm for communication in noisy environments,” *European Transactions on Telecommunications*, vol. 15, pp. 33–38, January/February 2004.

[38] F. Beritelli, S. Casale, and S. Serrano, “Adaptive v/uv speech detection based on acoustic noise estimation and classification,” *Electronic Letters*, vol. 43, pp. 249–251, February 2007.

[39] P. T. Brady, “A model for generating on-off speech patterns in two-way conversation,” *Bell Syst. Tech. J.*, pp. 2445–2472, September 1969.

[40] ETSI, “Gsm 06.94, digital cellular telecommunication system (phase 2+); voice activity detector (vad) for adaptive multi rate (amr) speech traffic channels; general description,” *Tech. Rep. V. 7.0.0*, February 1999.

[41] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using gaussian mixture speaker models,” *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72–83, 1995.

[42] “ITU-T Recommendation P. 59: Artificial conversational speech,” March 1993.

[43] F. Beritelli, S. Casale, and A. Cavallaro, “A robust voice activity detector for wireless communications using soft computing,” *IEEE J. Select. Areas Commun*, vol. 16, p. 18181829, December 1998.

[44] F. Beritelli, S. Casale, G. Ruggeri, and S. Serrano, “Performance evaluation and comparison of g.729/amr/fuzzy voice activity detectors,” *IEEE Signal Processing Letters*, vol. 9, pp. 85–88, March 2002.

[45] L. Couvreur and M. Laniray, “Automatic noise recognition in urban environments based on artificial neural networks and hidden markov models,” in *Proceedings of INTERNOISE*, 2004.

[46] F. Beritelli, S. Casale, A. Russo, and S. Serrano, “A speech recognition system based on dynamic characterization of background noise,” in *Proceedings of the 2006 IEEE International Symposium on Signal Processing and Information Technology*, pp. 914–919, 2006.

[47] T. B. Lee, *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its inventor*. Harper, 1999.

[48] A. Spadaccini, “Biometric4net web site and code repository: http://code.google.com/p/biometric4net,” 2011.

[49] World Wide Web Consortium (W3C), “HTML living standard - http://www.whatwg.org/specs/web-apps/current-work.”

[50] World Wide Web Consortium (W3C), “9.2 obtaining local multimedia content: http://www.whatwg.org/specs/web-apps/current-work/#obtaining-local-multimedia-content.”

[51] “Adobe Open Source - Flex 4 SDK: http://opensource.adobe.com/wiki/display/flexsdk/Download+Flex+4.”

[52] “Red5 - the open source media server: http://www.red5.org.”

[53] “erlyvideo streaming server: http://erlyvideo.org,.”

[54] “rtmplite: Flash RTMP server in python: http://code.google.com/p/rtmplite.”

[55] N. Poh and S. Bengio, “Can chimeric persons be used in multimodal biometric authentication experiments?,” in *Proceedings of the 2nd Joint AMI/PASCAL/IM2/M4 Workshop on Multimodal Interaction and related Machine Learning Algorithms (MLMI)*, pp. 11–13, 2005.

[56] D. Petrovska-Delacrtaz, G. Chollet, and B. D. (Eds.), *Guide to Biometric Reference Systems and Performance Evaluation*. Springer, 2009.

[57] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. Wiley, 1997.

[58] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, R. Preti, G. Pouchoulin, N. Evans, B. Fauve, and J. Mason, “Alize/Spkdet: a state-of-the-art open source software for speaker recognition.”

[59] B. Fauve, D. Matrouf, N. Scheffer, J.-F. Bonastre, and J. Mason, “State-of-the-Art Performance in Text-Independent Speaker verification through open-source software,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, Issue 7, pp. 1960–1968, 2007.

[60] The ELISA consortium, “The elisa consortium, the elisa systems for the nist99 evaluation in speaker detection and tracking,” *Digital Signal Processing*, vol. 10, 2000.