PhD Dissertation

FUSING PROSODIC AND ACOUSTIC INFORMATION FOR SPEAKER RECOGNITION

Mireia Farrús i Cabeceran

Thesis advisor: **F. Javier Hernando Pericás**

TALP Research Center, Speech Processing Group
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya

Barcelona, July 2008

Abstract

Automatic speaker recognition is the use of a machine to identify an individual from a spoken sentence. Recently, this technology has been undergone an increasing use in applications such as access control, transaction authentication, law enforcement, forensics, and system customisation, among others.

One of the central questions addressed by this field is what is it in the speech signal that conveys speaker identity. Traditionally, automatic speaker recognition systems have relied mostly on short-term features related to the spectrum of the voice. However, human speaker recognition relies on other sources of information; therefore, there is reason to believe that these sources can play also an important role in the automatic speaker recognition task, adding complementary knowledge to the traditional spectrum-based recognition systems and thus improving their accuracy.

The main objective of this thesis is to add prosodic information to a traditional spectral system in order to improve its performance. To this end, several characteristics related to human speech prosody —which is conveyed through intonation, rhythm and stress— are selected and combined with the existing spectral features. Furthermore, this thesis also focuses on the use of additional acoustic features —namely jitter and shimmer— to improve the performance of the proposed spectral-prosodic verification system. Both features are related to the shape and dimension of the vocal tract, and they have been largely used to detect voice pathologies.

Since almost all the above-mentioned applications can be used in a multimodal environment, this thesis also aims to combine the voice features used in the speaker recognition system together with other biometric identifiers —face— in order to improve the global performance. To this end, several normalisation and fusion techniques are used, and the final fusion results are improved by applying different fusion strategies based on sequences of several steps. Furthermore, multimodal fusion is also improved by applying a histogram equalisation to the unimodal score distributions as a normalisation technique.

On the other hand, it is well know that humans are able to identify others from voice even when their voices are disguised. The question arises as to how vulnerable automatic speaker recognition systems are against different voice disguises, such as human imitation or artificial voice conversion, which are potential threats to security systems that rely on automatic speaker recognition. The last part of this thesis consists of with an analysis of the robustness of such systems against human voice imitations and synthetic converted voices, and the influence of foreign accents and dialects —as a sort of imitation— in auditory speaker recognition.

Resum

El reconeixement automàtic del locutor és la utilització d'una màquina per identificar un individu a partir de d'un missatge parlat. Recentment, aquesta tecnologia ha experimentat un increment en l'ús de diverses aplicacions com el control d'accés, l'autenticació de transaccions, la cooperació amb la justícia, l'analítica forense o la personalització de serveis, entre d'altres.

Una de les qüestions centrals que es tracten en aquest camp és el fet de saber quina part del senyal de veu conté informació del locutor. Tradicionalment, els sistemes de reconeixement automàtic del locutor s'han basat principalment en característiques relacionades amb l'espectre de la veu. No obstant, els humans utilitzen altres fonts d'informació per reconèixer locutors, de manera que hi ha motius per pensar que aquestes fonts poden tenir un paper important en la tasca de reconeixement automàtic del locutor, aportar coneixement complementari als sistemes de tradicionals basats en l'espectre de la veu i millorar-ne la precisió.

L'objectiu principal d'aquesta tesi és incorporar informació prosòdica a un sistema espectral tradicional per tal de millorar-ne el funcionament. Amb aquesta finalitat, diverses característiques relacionades amb la prosòdia —constituïda per elements d'entonació, ritme i accent— es seleccionen i es combinen amb les característiques espectrals existents. A més a més, la tesi també se centra en la utilització de característiques acústiques addicionals —a saber, *jitter* i *shimmer*— per millorar el funcionament del sistema de verificació espectral-prosòdic proposat. Totes dues característiques estan relacionades amb la forma i la dimensió del tracte vocal, i s'han utilitzat en gran part per detectar patologies de la veu.

La majoria d'aplicacions que s'han esmentat abans es poden utilitzar en un entorn multimodal; per aquest motiu, les característiques de veu utilitzades en el sistema de reconeixement del locutor també es combinen amb altres identificadors biomètrics —concretament, la cara— per tal de millorar el funcionament global del sistema. Amb aquest objectiu, s'utilitzen diverses tècniques de normalització i de fusió, i els resultats de la fusió final es milloren aplicant diferents estratègies de fusió basades en seqüències de passos. A més a més, la fusió multimodal també es millora aplicant una equalització d'histogrames com a tècnica de normalització a les distribucions de puntuacions unimodals.

Per altra banda, és sabut que els humans poden identificar els altres a partir de la veu fins i tot quan aquestes veus estan alterades d'alguna manera. La qüestió rau en quina mesura els sistemes automàtics de reconeixement del locutor són vulnerables a les diferents alteracions de la veu, com ara la imitació humana o la conversió artificial. L'última part de la tesi consisteix en una anàlisi de la robustesa d'aquests sistemes a les imitacions de veu humanes i a les veus convertides sintèticament, i de la influència dels accents estrangers —com a tipus d'imitació— en el reconeixement auditiu del locutor.

Agraïments

Confesso que he estat molt temptada de deixar aquesta pàgina en blanc per un sol motiu: no sabia com començar. Però al final he fet un balanç de motius: en tinc moltíssims per donar les gràcies i un de sol per no fer-ho. I com que d'alguna manera he de començar, he decidit optar per l'ordre cronòlogic. Així doncs...

... gràcies, Núria Bel. Et pertoca ser la primera pel post-it groc on vas escriure

```
Javier Hernando
     UPC
```

i per aconsellar-me que fes la tesi amb ell. Va ser, sens dubte, un gran consell. Gràcies, Ramon, per escoltar la meva història sobre el post-it groc i ajudar-me en els primers passos. I gràcies, Javier, per haver acceptat ser el meu director de tesi tal com deia el post-it groc, i pel teu suport durant aquests anys.

Després del post-it groc va venir l'abisme: què és això del cepstrum i què hi faig aquí. Confesso també que el primer dia vaig estar temptadíssima d'anar-me'n a casa i no tornar; però els altres doctorands em van donar molts motius per no fer-ho. Gràcies Jan, Pere, Mònica, Pablo, Jordi, Josep Maria, Marta Casar, Marta Costa-Jussà (germaneta!), Andrey, Cristian, Enric, Martí i molts més. Per fer-me reflexionar sobre el caràcter acumulatiu de les monedes i l'efecte afrodisíac dels microones, pels avions de paper, les magdalenes de la FIB i moltíssimes coses més. Gràcies, Xavi, per ser el meu informàtic incondicional durant el primer any de l'abisme. Gràcies també al Carlos Nistal per escoltar sempre amb un somriure les meves peripècies informàtiques, al Pascual per ajudar-me pacientment amb els experiments sempre que l'hi he demanat, i als que heu fet possible l'últim capítol de la tesi: Adrià, Jordi Ventura, Queco, Cesc i Víctor Polo.

Als que vau fer que l'hivern suec a vint-i-tres graus sota zero fos una mica més càlid: Kirk, Tomas (per trobar-me un sostre), Urban, Linnea, Misuzu, Fredrik (avui tampoc et preguntaré per les plantes), Erik, Mattias, Leila, Thierry, Elisabeth i Ingmarie, tak so mycket. I gràcies als que, molt lluny d'aquí a l'altra banda del món —sense ser barrufets— vau fer que em sentís com a casa; gràcies, Michael, per fer possible l'estada a Austràlia, gràcies Peter i Del per fer-me de pares durant sis mesos, gràcies Ian i Elizabeth per portar-nos a platges amb taurons i meduses mortals i per la vostra companyia; gràcies Girija, Zoe, Rod, Claudia, Diane, Dave i Ford Laser, cangurs, cacatues, pelicans, ibis i roselles. Us tinc a tots en un trosset del meu cor, i el meu enyorament és tan gran com el meu agraïment. I gràcies, Jan, per compartir-ho amb mi i pel primer Ambalindum. Espero que n'hi hagi molts més.

Finalment, gràcies família, pel vostre suport incondicional i perquè no entreu en cap ordre cronològic; hi sou a tot arreu, al principi i al final, tant en el temps com en l'espai. I a més a més, tinc la sort de saber que sempre serà així. Moltes gràcies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speech processing by computer is a field that encompasses, among others, computer science, linguistics and speech communication. This multidisciplinary field is normally divided into three distinct subfields: speech synthesis, speech recognition and speaker classification.

Speaker classification is mainly concerned with extracting information about individuals from their speech. This includes guessing whether the speaker is male or female, adult or child, their emotional state (Hecker et al., 1968; Williams and Stevens, 1972), the language the person is speaking, the education level and social rank, or even the speaker's nationality (Baldwin and French, 1990).

A speaker's identity can also be determined by the speech signal. This task is known as speaker recognition, which is one of the most widely investigated subareas in speaker classification (Kersta, 1962; Stevens et al., 1968; Bolt et al., 1969; Atal, 1972; Doddington, 1985; Soong et al., 1985; Furui, 1996; Campbell, 1997; Klevans and Rodman, 1997), and the focus of this thesis.

Next, an overview of the most important applications in speaker recognition is described; the motivation and main objectives of this thesis are presented in section 1.2 and, finally, the structure and contents of the thesis are outlined in section 1.3.

1.1 Speaker recognition applications

Automatic speaker recognition is the use of a machine to identify an individual from a spoken sentence. In a human speaker recognition process, the better one knows a person, the easier it is to identify others by their speech. Like humans, automatic speaker recognition systems need a training period to learn what speech is like.

Speaker and speech recognition use similar speech signal processing techniques. However, speech recognition —if it is to be speaker independent—, focuses on those aspects of the speech signal carrying more linguistic information, whereas speaker recognition is based on those idiosyncratic speech features that characterise an individual.

Research on speaker recognition began in the 1960s when scientists attempted to use the speech spectrogram as a tool for speaker recognition (Kersta, 1962; Stevens et al., 1968; Bolt et al., 1969; Tosi et al., 1972). Computer technology was, at that time, not advanced enough to complement the manual work of phoneticians interpreting the spectrograms.

According to Reynolds et al. (2002), there are two main factors that make human voice a compelling characteristic to recognise people: in the first place, speech is a natural signal to produce that is not considered threatening by users to provide. Second, the telephone system provides currently a ubiquitous, familiar network of sensors for obtaining and delivering the speech signal. Therefore, nowadays, there is a wide variety and continued growth of applications based on speaker recognition technologies. The most common areas where these applications can be found are listed next.

Access control Applications related to secure access to physical and electronic site are probably the most common ones. These applications have the advantage that, unlike personal passwords and keys, voices cannot be stolen. However, they can be copied by using, for instance, recording devices. In order to protect security systems from this risk, the speaker wishing to access to a secure place is usually asked to pronounce a specific text. In this case, both speaker and the linguistic content of the speech are taken into account.

Transaction authentication In addition to access control, higher levels of verification may be needed for telephone banking, for example, in order to achieve more secure transactions. Some recent applications are also focused on user authentication tasks for remote electronic purchases and both fixed telephone and mobile shopping.

Law enforcement and forensics Law enforcement includes several applications such as home-parole monitoring —where parolees are called in order to check that they are staying at home— prison call monitoring, border controls, etc. Forensic applications are highly related to law enforcement applications, especially those concerning location of missing people and criminal identification.

Speech data management The use of speaker recognition is also incipient in several applications such as voice mail browsing or intelligent answering machines, where incoming voices are labeled with the speaker name. Speech data management can also be found in smart rooms to automatically track who said what, for example, in a boardroom meeting.

Personalisation More and more, a wide variety of devices and smart systems can be found to organise and facilitate our daily life. Presumably, those devices controlled by voice will perform better with a good personal customisation. Furthermore, there is also an incipient interest in using speaker characterisation in order to provide personal information to be used in advertisements or other services.

Speaker recognition applications can be classified into *text-dependent* and *text-independent* applications. In the former, the speaker is required to pronounce the same text in both training and testing steps. In the latter, the applications do not rely on any specific spoken message. In this case, it is more difficult to achieve a good accuracy; however, the task becomes more flexible.

Almost all the applications listed above can be used in a multimodal environment; i.e. the speaker recognition task can be combined with recognition technologies involving other modalities. For example, fingerprints or iris scan can be used to control access, and facial features can be used, for instance, in smart meeting rooms and customisation of services.

1.2 Motivation and objectives

One of the central questions addressed by automatic speaker recognition research is what is it in the speech signal that conveys speaker identity. We, humans, rely on several distinct types or levels of information contained in the speech signal in order to identify people from voice alone (Schmidt-Nielsen and Crystal, 2000). These can be related to diverse aspects such as the voice timbre, a characteristic laugh or specific words repeatedly used.

Traditionally, these levels of information have been roughly hierarchised from low-level to high-level information. Low-level information has been mainly associated to those features related to the physical traits of the vocal apparatus —such as the voice timbre—, whereas high-level is mostly associated to those features depending on the learned habits and style —such as a particular word usage.

Automatic speaker recognition systems have relied mostly on low-level characteristics by using short-term features related to the spectrum of the voice. However, other levels of information play an important role in the human recognition process, which means that they convey useful speaker information. Therefore, there is reason to believe, as recent studies have demonstrated (Carey et al., 1996; Sönmez et al., 1998; Doddington, 2001; Andrews et al., 2002; Bartkova et al., 2002; Weber et al., 2002; Peskin et al., 2003; Reynolds et al., 2003), that these levels can add complementary knowledge to the traditional spectrum-based recognition systems, improving their accuracy. Moreover, they appear to be more robust to acoustic degradations from channel and noise effects (Atal, 1972; Carey et al., 1996).

The main objective of this thesis is to add prosodic information to the traditional spectral systems in order to improve their performance. The idea is to find and select appropriate characteristics related to the human speech prosody —which is conveyed through intonation, rhythm and stress— and to combine them with the existing and widely used spectral features. Such prosodic characteristics include parameters related to the fundamental frequency in order to capture the intonation contour, and other parameters such as the length of the speech segments, as well as the rate and duration of pauses, in order to represent the speaker speech tempo.

The use of prosodic features for speaker recognition requires a set of characteristics that should be taken into consideration. Just as systems relying on short-term spectral features can achieve a good performance with a reasonable low amount of training and testing data, prosodic features need much more data in order to capture the speaker prosodic style. Moreover, prosodic characteristics related to a specific speaker are especially manifested when speaking in spontaneous style. For this reason, a conversational speech database is required in order to obtain a reasonable reliable analysis. In this thesis, the Switchboard-I database will be used in the proposed prosodic recognition system in order to ensure both dimension and spontaneity requirements.

Apart from the above-mentioned prosodic features, there are many more characteristics that may provide complementary information and should be of great value for the speaker recognition task. This thesis focuses also on the use of two additional acoustic features for speaker verification systems: *jitter* and *shimmer*. These features are, somehow, related to the shape and dimension of the vocal tract, and the way how the speech is produced. Specifically, jitter and shimmer have been largely used to detect voice pathologies and to identify the age and gender of the speakers, which leads to think that they can be of usefulness in the speaker recognition task. The improvement of traditional systems by using these additional features is also an objective of this thesis.

These complementary features are especially useful when combined within them and with the short-term spectral parameters; when used alone, the performance of individual parameters is normally too worse to be used to recognise individuals in a reliable manner. In the same way, the results achieved in a person recognition task can be improved by using more than a single biometric modality. To this end, multimodal biometric fusion combines the information from multiple modalities. As well as the use of complementary speech features for speaker recognition, the objectives of this thesis include the combination of such features with face related parameters in order to improve the global performance of the system. This combination of features —both within the same modality and between different modalities— involves the use of normalisation and fusion techniques. Therefore, the above-mentioned tasks include other objectives: to implement existing normalisation and fusion techniques, to find those techniques with which the best performance is obtained, and to try to improve the corresponding normalisation and fusion steps with new methods and strategies.

Nevertheless, although the performance of biometric recognition systems can be highly improved when using different biometric identifiers, these systems are still exposed to several external threats. Voice imitation and other types of disguise are, for instance, potential threats to security systems that rely on automatic speaker recognition. It is well-known that human speaker recognition can be quite reliable in small sets of people and specially when the listeners are familiar with the speakers. Even more, humans are able to identify others from voice even when their voices are disguised (Reich, 1981). However, the question arises as to how vulnerable automatic speaker recognition systems are against different voice disguises, such as human imitation or artificial voice conversion. The last objective of this thesis is to test the robustness of speaker recognition systems against human and synthetic voice imitation.

1.3 Thesis overview

This thesis is structured as follows. The state of the art in speaker recognition is reviewed in chapter 2. This chapter includes several parts. The first one is a brief description o biometrics, focusing on the general performance of automatic biometric recognition systems and their areas of application. The second one deals with several aspects about speaker recognition: the differences between automatic and human recognition, the speaker information that can be found in the speech signal, and some of the most common automatic speaker recognition techniques. Next, the role of prosody in speaker recognition is reviewed, followed by the description of fusion techniques used to combine different biometric modalities. Finally, the chapter contains a brief state of the art in voice imitation and conversion in speaker recognition, focusing on the features and techniques normally used and their robustness against automatic recognition systems.

Chapters 3, 4 and 5 contain the experimental part of the thesis. In chapter 3, a prosodic speaker recognition system is proposed in order to improve a baseline spectral system. Partly inspired by the works of Shriberg et al. (2000) and Peskin et al. (2003), some prosodic features are captured and used in combination with spectral features. Several normalisation and fusion techniques are applied to fuse the prosodic features, and the same techniques are then used to combine both prosodic and spectral systems. Furthermore, other acoustic features —namely jitter and shimmer, which have been widely used to detect voice pathologies— are introduced in order to improve the performance of both prosodic and spectral systems.

Chapter 4 focuses mainly on the importance of normalisation and fusion techniques in multimodal biometric recognition systems. The prosodic and spectral features used in the previous chapter are combined with facial parameters, obtaining a multimodal person recognition system based on speech and face parameters. Apart from normalisation and fusion methods used in chapter 3, fusion by support vector machines and a histogram equalisation as a normalisation technique is applied to the biometric scores. Furthermore, several fusion strategies based on the use of different fusion steps are proposed.

The experimental part of this thesis finishes in chapter 5 with an analysis of the vulnerability of the systems against imitated and converted voices. First, some experiments are performed in order to test the influence of foreign accents and dialects —as a sort of imitation— in auditory speaker recognition. Second, the voices of two well-known professional imitators trying to impersonate several well-known politicians are used to analyse the behaviour of some selected prosodic and acoustic features in the imitated voices. At the end of the chapter, converted voices are also used to test the robustness of speaker verification systems.

Finally, conclusions of the experiments and some proposed guidelines for future work are presented in chapter 6.

Chapter 2

State of the Art in Speaker Recognition

Automatic speaker recognition is the use of a machine to recognise an individual from a spoken sentence. Thus, a speaker recognition process is always concerned with extracting information about individuals from their speech.

Speech signal processing extracts features from the speech signal, which relate to the manner of sound generation in the larynx (*source*) on the one hand and to the acoustic filtering of the speech sounds in the vocal and nasal tracts (*filter*) on the other. Early automatic speaker recognition systems tended to use solely the filter parameters, which relate —in a complex way— to the physiology of the vocal tract and to the learnt articulatory configurations that shape the specific speech sounds (Rabiner and Juang, 1993). These features have also been being used in multimodal person recognition systems, where two or more human traits like voice, face, fingerprints, iris, hand geometry, etc. are involved in the recognition process.

More recently, some speaker recognition systems have begun to use also the source parameters, which relate mainly to the fundamental frequency and power (or perceived pitch and loudness) of the speech sounds and, in turn, to the prosody of spoken phrases (Peskin et al., 2003; Reynolds et al., 2003). Generally, systems that use both source and filter parameters perform better than systems that just use source parameters, when systems are evaluated by means of generic background models and without impostors who employ intentional voice mimicking techniques.

Some recent studies have tested the vulnerability of automatic speaker recognition systems to intentional voice mimicking (Lau et al., 2004, 2005). Such vulnerability is of particular concern where automatic speaker recognition is used to control client access in applications such as telephone banking or other financial services. Where a speaker recognition system uses both source and filter parameters, the question arises whether either the source or the filter parameters are more vulnerable to intentional mimicking. Moreover, current speech synthesis technology is more and more able to approximate the voice of a specific target speaker —as a sort of voice mimicking—, which raises the doubt about how reliable are security systems against such synthetic voices.

The state-of-the-art of these main topics is covered in this chapter. Section 2.1 is focused on the most used features in biometric recognition, the performance of automatic biometric recognition systems and their main applications. Section 2.2 contains a brief description of the speech production and the state of the art in current speaker recognition technologies. A more detailed report about the role of prosody in speaker recognition task is presented in section 2.3. The following section (2.4) is about fusion of different modalities, focusing on the most commonly used normalisation and fusion techniques. Finally, section 2.5 is focused on the importance of voice disguising in speaker recognition technologies.

2.1 Biometrics

More than a century ago, Alphonse Bertillon first conceived and then industriously practiced the idea of using body measurements for solving crimes (Jain et al., 1994). Nowadays, biometric recognition refers to the use of distinctive characteristics to recognise individuals (Maltoni et al., 2003; Bolle et al., 2004). These distinguishing characteristics, called biometric identifiers —or simply biometrics—, are usually classified into physiological or behavioural characteristics.

Physiological biometrics, like fingerprints, face, hand geometry, retina or iris, are physical characteristics that can be measured at some point of time. On the other hand, behavioural biometrics like, like signature, voice or gait, consist of the way some action is carried out and extend over time. Unlike physiological biometrics, behavioural biometrics are learned or acquired over time and they can be easily and deliberately changed (Bolle et al., 2004).

Nevertheless, although the classification of biometric identifiers into behavioural and physiological characteristics seems to be clear a priori, mostly all biometric identifiers are, in some way, a combination of both characteristics. Behavioural biometrics are related to movement or dynamics, but they are highly dependent on the physiological structure of the individual; voice and gait, for example, depend on the anatomical structure of the human vocal mechanism and the legs, respectively (Brand et al., 2001).

At this point, the question arises as to which human characteristics can be used as biometric identifiers. Maltoni et al. (2003) reported that

> any human physiological and/or behavioural characteristic can be used as a biometric identifier to recognise a person as long as it satisfies the following requirements:

- **Universality**: each person should have the biometric.
- **Distinctiveness**: any two individuals should be sufficiently different in terms of their biometric identifiers.
- **Permanence**: the biometric should be sufficiently invariant over a period of time.
- **Collectability**: the biometric can be measured quantitatively.

However, in practical biometric systems, some other issues like the system accuracy, intrusiveness (undesirable contact with the subject in order to acquire the biometric data) or acceptability (to which extent people are willing to accept a particular biometric identifier in their daily lives) should be also considered.

The choice of which biometric identifiers should be used depends on the application, since each biometric has its advantages and disadvantages. Next, the most common biometrics are briefly described, focusing on their main characteristics, strengths and weaknesses, and applications. These characteristics are summarised and compared in Table 2.1, according to (Maltoni et al., 2003).

2.1.1 Biometric identifiers

Face Face is one of the most acceptable biometrics, since it is one of the most common methods of recognition that humans use in their visual interactions (Maltoni et al., 2003). In addition, unlike other —more reliable— methods of biometric personal identification like fingerprints, retinal or iris scan, face recognition is non-intrusive, and does not rely in the cooperation of the participants (Zhao et al., 2003).

Given still or video images of a scene, the aim is to recognise one or more individuals in the scene using a stored database of faces (Zhao et al., 2003). The main challenge is to develop face recognition techniques that can tolerate the effects of aging, facial expressions, slight variations in the imaging environment, and variations in the pose of the face with respect to the camera (Maltoni et al., 2003; Jain et al., 1994).

Fingerprints Fingerprints are the most commonly used biometrics due to their well-known distinctiveness and permanence properties over time. In 1893, the Home Ministry Office (UK) accepted that no individuals have the same fingerprints. Since then, this biometric identifier has been widely used in forensic applications for over one hundred years, and automatic fingerprint identification systems were introduced as one of the first applications of machine pattern recognition almost fifty years ago (Maltoni et al., 2003).

Fingerprints are characterised by medium universality, acceptability and collectability, and a high accuracy. Even so, fingerprint recognition is still a challenging and an important pattern recognition problem (Maltoni et al., 2003; Jain et al., 1994).

Voice Voice is not expected to have enough distinctiveness to allow the recognition of an individual from a large database of speakers. Moreover, it is characterised by three other disadvantages: first, a speech signal can be degraded in quality by the microphone or the transmission channel; second, voice can be affected by a person's health, stress or emotions; and finally, it has been shown that some people are extraordinarily skilled in mimicking others' voices (Maltoni et al., 2003; Bolle et al., 2004).

However, voice is a non-intrusive biometric with a high acceptability. Moreover, it is nowadays the only feasible biometric identifier in applications requiring person recognition over a telephone system, which provides a ubiquitous and familiar network of sensors for obtaining and delivering the speech signal (Maltoni et al., 2003; Bimbot et al., 2004).

Iris Iris recognition technology is highly accurate and fast. The visual texture of the human iris is determined by the chaotic morphogenetic processes during embryonic development and is believed to have a high distinctiveness for each person and each eye. Universality and permanence over time are also high, and its weaknesses rely on a medium collectability and a low acceptability, since capturing an iris image is rather intrusive and requires cooperation from the user (Maltoni et al., 2003; Jain et al., 1994). Iris recognition applications have been recently used in access control to hospitals and international frontiers.

Hand and finger geometry Hand and finger geometry recognition are characterised by a relatively high permanence, although such features related to a human hand are not very distinctiveness to individuals. They are also characterised by a medium acceptability, since the acquisition systems involve the cooperation of the user and they are perceived as quite intrusive methods. Their strength relies on the collectability, since the representational requirements of the hand are very small. Even so, finger geometry systems are sometimes preferred due to their more compact size (Maltoni et al., 2003; Jain et al., 1994).

Signature Each individual seems to have a characteristic way of signing, and signature has a high acceptability in many legal and commercial transactions as a verification method. Furthermore, the collectability is also very high. Its weaknesses in biometric recognition are the low universality, distinctiveness and permanence, which entails a low accuracy of the systems. Signatures change over a period of time or even between successive impressions of the same user. Furthermore, signatures can be easily be forged by professionals (Maltoni et al., 2003).

Ear The shape of the ear and the structure of the cartilaginous tissue of the pinna are not expected to be unique to an individual, but they are distinctive enough to be used in biometric recognition. Furthermore, they are characterised by a high user acceptability and permanence over time (Jain et al., 1994).

DNA Deoxyribonucleic Acid (DNA) is the one-dimensional ultimate unique code for one's individually, used mostly in forensic applications for person recognition (Maltoni et al., 2003; Jain et al., 1994). Its high distinctiveness and permanence over time entail a high performance of the system based on this biometric. However, DNA as a biometric identifier is characterised by a low collectability and a low acceptability, since it is a very intrusive acquisition method.

Gait The way a person walks doesn't seem to be very distinctive, but it is supposed to be peculiar enough to permit person verification in some low-security applications. Gait is not permanent over time, specially over a large time period. However, since acquisition of gait is similar to acquiring facial pictures, it is a very acceptable and non-intrusive biometric. Furthermore, it is characterised by a high degree of collectability (Maltoni et al., 2003; Jain et al., 1994).

Retinal scan As the iris visual texture, the retinal vasculature seems to be a characteristic of each individual and each eye. It is claimed to be the most difficult to forge biometric. However, the image acquisition is quite intrusive and requires a large cooperation and effort of the user; therefore, this method is characterised by a low public acceptability (Maltoni et al., 2003; Jain et al., 1994).

Odor A human body emits a chemical component that is distinctive to a particular individual. A part of this high distinctiveness, odor seems to be permanent over time and its acquisition is quite accepted by users. However, odor data is not easy to collect, and although its characteristic distinctiveness and permanence, the performance of the systems is still a challenge in biometric recognition (Maltoni et al., 2003; Jain et al., 1994).

Keystroke dynamics Keystroke dynamics refers to the characteristic way each individual types on a keyboard. So far, this is probably the least used biometric due to its large list of weaknesses: keystroke dynamics is not very distinctive nor permanent over time. It is also one of the least universal biometrics, since not everybody uses keyboards in the daily life. All these characteristics entail a low performance of the systems. However, the acquisition of keystrokes data is a non-intrusive method, so that it is rather acceptable within the users (Maltoni et al., 2003; Jain et al., 1994).

Biometric identifier	Universality	Distinctiveness	Permanence	Collectability
Face	H	L	M	H
Fingerprints	M	H	H	M
Voice	M	L	L	M
Iris	H	H	H	M
Hand and finger geometry	M	M	M	H
Signature	L	L	L	H
Ear	M	M	H	M
DNA	H	H	H	L
Gait	M	L	L	H
Retinal scan	H	H	M	L
Odor	H	H	H	L
Keystroke dynamics	L	L	L	M

Table 2.1: Comparison of several biometric identifiers. Data based on the perception of Maltoni et al. (2003). H, M and L denote High, Medium and Low, respectively.

2.1.2 Automatic biometric recognition systems

Architecture of a recognition system

A typical biometric recognition system consists mainly of two phases: the training —or enrolment— phase and the testing —or recognition— phase. In the training phase, biometric measurements from the users are captured by biometric sensors or readers. Next, relevant information is extracted in order to build a user model, which will be stored in a database.

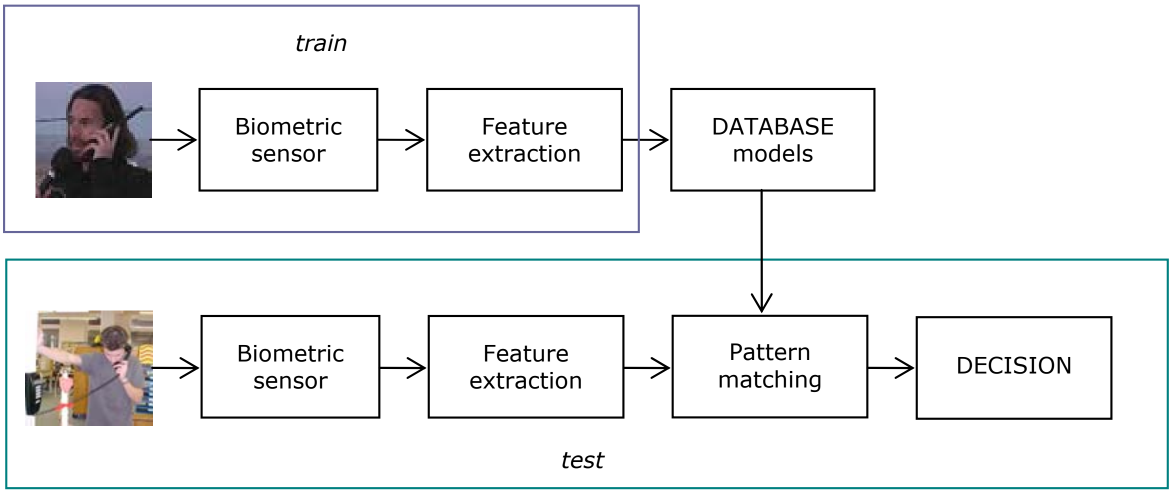


Figure 2.1: Architecture of a typical biometric recognition system.

In the recognition phase, biometric readers are also used to capture biometric information of the user to be recognised. Relevant information is then extracted in the feature extraction step. Next, this information is compared with the stored user models of the database, computing the degree of similarity —or *score*—, which will be used the determine the likelihood that the user corresponds to one of users whose model is stored in the database. In the end, a decision will be taken based on the computed similarity scores.

Biometric identification and verification

Depending on the type of application used, an automatic biometric recognition system can run in two modes: identification and verification (Campbell, 1997).

In a person **identification** mode, the aim is to determine which speaker, in a set of users whose models are stored in the database, matches an unknown user. The set of users can be a closed or an open set. In a closed set identification, it is assumed that the unknown user matches one of the known users of the database, and the biometric recognition system simply choose the known user that mostly matches the unknown user. In an open set identification, the unknown user may or may not be in the set of known users. In this case, it is like performing a closed set identification followed by a verification process to ensure that the unknown user is close enough to the chosen known user of the database to be identified as such. Figure 2.2 shows a diagram of an automatic identification system.

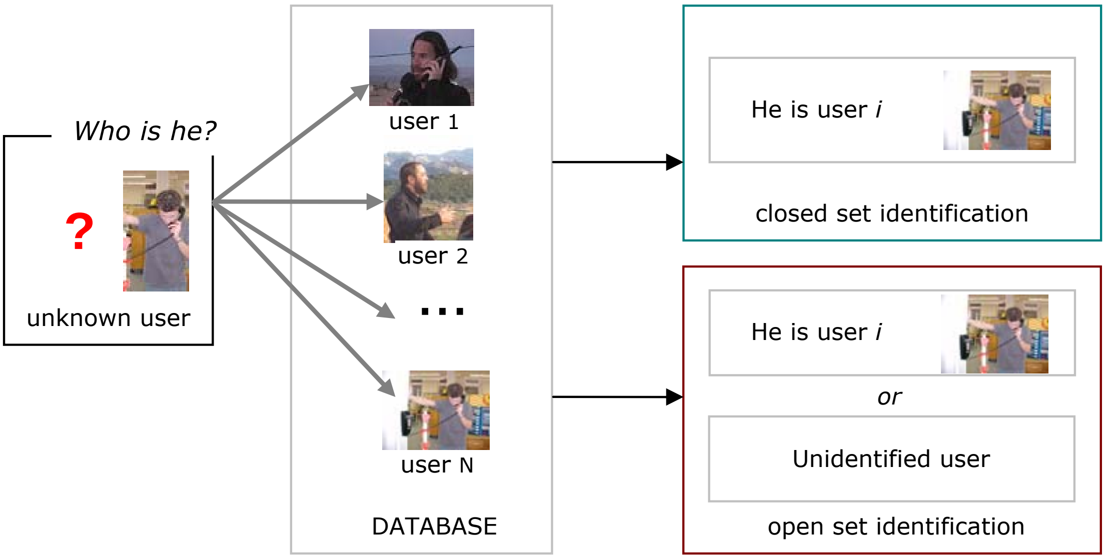


Figure 2.2: Automatic identification system.

On the other hand, the aim of a system running in the **verification** mode (Figure 2.3) is to determine whether users are who they claim to be. Applications of this recognition mode are mainly related to access restriction in secure areas. In verification systems, a user is claiming an identity. A model corresponding to that identity must be stored in the database, which must contain an impostor model as well. Then, the biometric features of the claimed user will be compared to the model of the claimed identity and to the impostor model. If a user seems to be closer to the claimed identity, they will be accepted as a known user (or client). Otherwise, the user will be rejected and treated as an impostor.

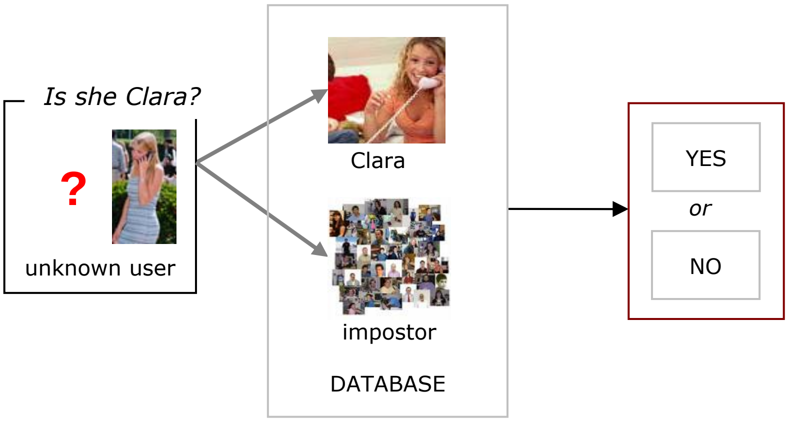


Figure 2.3: Automatic verification system.

Evaluation of a biometric verification system

After having computed a match score of similarity between the input user and the corresponding template stored in the database, a decision is taken whether the user must be accepted or rejected by the system. However, such decision can be both correct or not correct. If the decision is incorrect, two different types of error can occur (Bimbot et al., 2004):

- **False rejection (or non detection)**: the system rejects a valid identity claim.
- **False acceptance (or false alarm)**: the system accepts an identity claim from an impostor.

Both types of errors give rise to two types of error rates, which are commonly used to measure the performance of a system:

- **False rejection rate (FRR)**: percentage of incorrectly rejected clients.
- **False acceptance rate (FAR)**: percentage of incorrectly accepted impostors.

Either of the two types of errors can be reduced at the expense of an increase in the other, so that the trade-off between FRR and FAR depends on a decision threshold. In a real-world system, which is usually not perfect, FRR and FAR intersect at a certain point (see Figure 2.4). The value of FRR and FAR at this point is known as the *Equal Error Rate* (EER).

If the threshold is set to a low value, the system tends to accept most of the identity claims, giving few false rejection errors but many false acceptances. On the contrary, with a high threshold the system tends to reject most of the identity claims, giving rise to few false acceptance errors and a lot of false rejections.

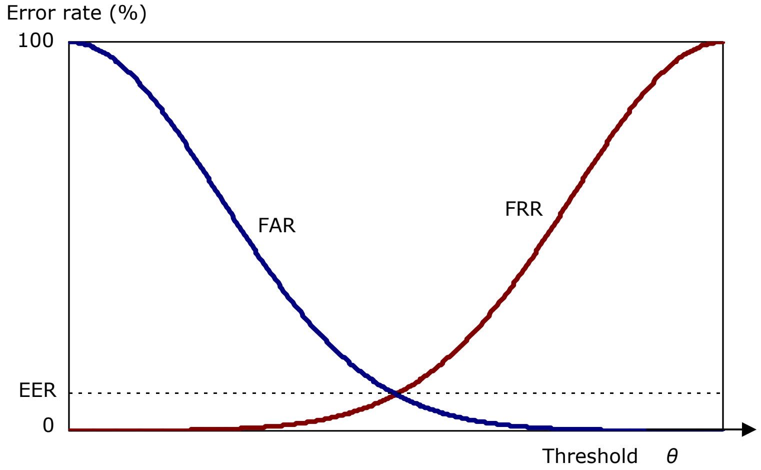


Figure 2.4: FRR and FAR as a function of a threshold θ. The intersection point determines the EER.

Therefore, when designing a biometric verification system, the decision threshold must be adjusted so that both errors are as low as possible, or one of the errors must be always below a certain threshold, if this property is required by a specific application. The trade-off between the two types of error rates is usually depicted in different ways. Two of the most common representations are the ROC and the DET curves.

The *Receiver Operating Characteristic* (ROC) curve plots the FRR versus the FAR (Bimbot et al., 2004). This curve is monotonous and decreasing, and the better the system is, the closer to the origin the curve will be. Another representation of the ROC curve is used sometimes by plotting the correct detection rate (instead of FRR) versus the false alarms (Duda, 2001).

It is also common to plot the error curve on a normal deviate scale. In this case, the curve is known as the *Detection Error Trade-offs* (DET) curve (Martin et al., 1997; Bimbot et al., 2004). In an hypothetical system whose client and impostor scores are Gaussians with the same variance, the DET curve is a linear curve where the slope equals -1, which becomes more easily readable and comparable to other DET curves. As in the ROC curve, better systems are closer to the origin. In a real system, the score distributions are not exactly Gaussians, but they are close enough to them to allow this representation.

Examples of both ROC and DET curves are plotted in Figure 2.5. The intersection of each curve with the diagonal dotted line indicates the value of the EER.

2.1.3 Application areas of biometric recognition

The need of recognising people in many daily life situations has given rise to many application relying on biometrics. This section reports some of the areas where this technology can be found: access restriction, electronic commerce transactions, personal service customisation, forensic applications and law enforcement.

The type of application will definitely determine the characteristics of each system, namely, which biometric identifiers are going to be used, the degree of security and restriction, the

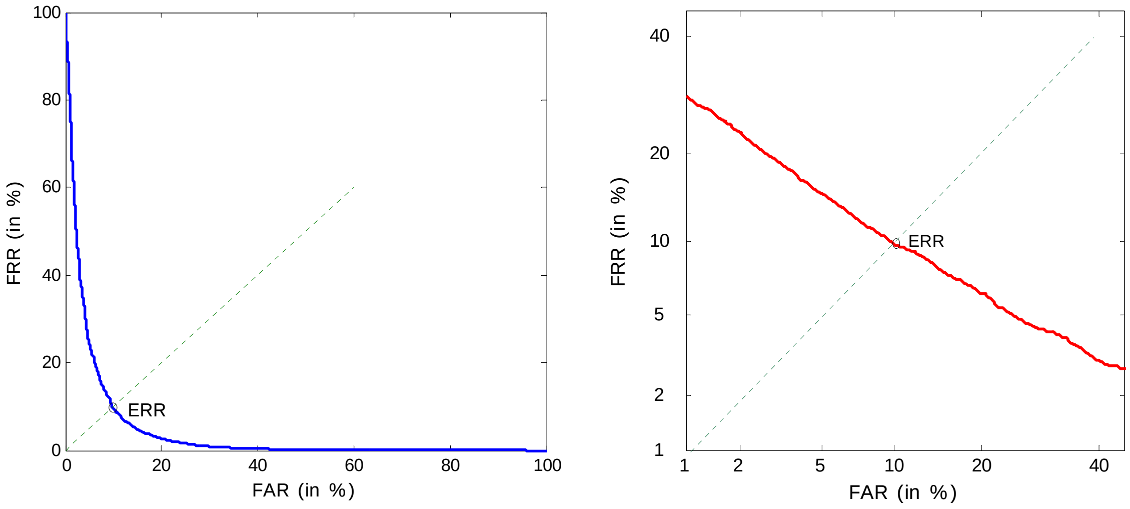


Figure 2.5: Examples of a ROC curve (a) and the corresponding DET curve (b).

environment or recording conditions, and the characteristics of the database, among others. The most common applications are briefly described next, and summarised in Table 2.2.

Access control

Access control and restriction is one of the areas in which biometric technologies have had the greatest impact. So far, access security has been normally restricted by using several objects or codes like keys, cards, passwords or personal identification numbers. Nevertheless, these personal objects and codes can be lost, stolen, or copied, so that biometric identifiers provide an alternative to access secured areas.

Transaction authentication

Online banking, trading and electronic purchasing transactions need verification methods to avoid frauds. Signature and pin numbers are currently the most used verification methods in electronic commerce. However, as in the access restriction applications, these methods can be easily copied by professionals. On the other hand, signature cannot be used in some applications such as telephone or remote credit card purchases. Therefore, other biometric identifiers will be even more used to increase the security in this kind of applications.

Service customisation

Normally, offices, cars, buildings, etc., contain devices and smart systems to facilitate our daily life. Domotics (or home automation), for example, consists in the use of systems and devices to provide security and comfort to the residents: light and climate control, remote use of home entertainment systems, etc. These applications will perform better with a good customisation of the systems, which are usually trained and controlled by biometric identifiers.

Law enforcement

Fingerprint identification was the first application of biometrics in law enforcement. Since biometrics is even more being used in daily life, it will also be increasingly used in law enforcement applications: locating missing people, identifying criminals, border controls, etc. Unlike security applications, which usually require that the recognition system respond within a short period of time, applications occurring in law enforcement may not have this constraint (Klevans and Rodman, 1997).

Forensics

Biometric forensic applications are highly related to law enforcement applications, especially those concerning the location of missing people and criminal identification. However, forensics has its own role in other issues not related to the law, such as mass disaster victim identification. In these cases, the use of biometric identifiers such as the teeth have been crucial to identify victims.

Application	Examples
Access control	Physical facilities Computer network and website
Transaction authentication	Online banking Electronic commerce
Service customisation	Intelligent answering machine Domotics
Law enforcement	Home parole
Forensics	Mass victim identification

Table 2.2: Application areas of biometric recognition.

2.2 Speaker recognition

This section deals with several aspects about speaker recognition: the differences between automatic and human recognition, the speaker information that can be found in the speech signal, and some of the most common automatic speaker recognition techniques.

2.2.1 Auditory and automatic speaker recognition

Human beings and computers recognise speech patterns in different ways (Ladefoged, 2001; Alexander et al., 2005). Humans use normally aural —hearing perception—, linguistic and other kinds of knowledge to identify others from their voices (Hollien, 2002; Rose, 2002); on the contrary, automatic speaker recognition systems rely on the use of acoustic properties extracted by computers (Rose, 2002).

Humans are able to recognise speakers even when their voices are disguised (Reich, 1981). This ability, however, is highly dependent on the familiarity of the listener with the speaker. It is well known that human beings are able to identify familiar speakers on the telephone after listening to a very short segment of speech, but they find more difficult to recognise the voices of less familiar speakers. Nevertheless, humans can be *trained* by listening repeatedly these less familiar voices in order to improve the recognition. Automatic speaker recognition systems can also be trained to learn the voice characteristics of a specific person. Like humans, the more training data, the better the speaker recognition will be.

Rose (2002) reports how the earlier history of forensic-phonetic activity was divided into the use of acoustic and auditory analysis. With respect to these two different approaches, three radically different positions were encountered:

- auditory analysis is sufficient on its own (Baldwin, 1979; Baldwin and French, 1990);
- auditory analysis is not necessary at all: it can be done with acoustics; and
- auditory analysis must be combined with other, i.e. acoustic, methods (Künzel, 1987; French, 1994; Künzel, 1995).

The third approach is the common method used nowadays in the forensic-phonetic context. However, auditory analysis cannot cover the use of a large number of speakers, which is one of the aims of the state-of-the-art automatic speaker recognition systems. Moreover, auditory analysis cannot be used neither in most of the real-time applications that need to identify many people in a very short-period of time; therefore, its use is normally restricted to some forensic applications.

2.2.2 Speaker information contained in the speech signal

Speech production

By means of spoken utterances, a person is able to communicate a message made up of sounds, words and sentences with a specific meaning. But besides carrying a meaning, the speech message can also provide information about the individuals themselves: the anatomy, physiology, age, linguistic characteristics or even emotional states (Dellwo et al., 2007).

The part of the human vocal mechanism responsible for the sound production —when driven by the brain— is the vocal tract. The elements of the human vocal tract were formerly associated only with vegetative functions —the tongue for mastication, the epiglottis and pharynx for swallowing, etc. These structures evolved, about 1.6 million years ago, in their ability to produce speech in order to enhance the human communication. This capability evolved in parallel with the peripheral hearing mechanism and the expansion in the neural circuitry for controlling muscles to produce sounds and for making sense of the incoming acoustic signals, giving rise to the beginning of human speech (Rose, 2002).

The human vocal system (Figure 2.6) is part of the human anatomy driven by an excitation source generated by airflow from the lungs. The air flows along the trachea and through the vocal folds, which open and close rapidly due to a combination of factors, producing a sort of vibration at the same time that the airflow is modulated. As the vocal folds open and close, jets of air flow through them; the frequency of these pulses determines the fundamental frequency and contributes to the perceived pitch of the produced sound.

The area between the vocal folds is called the glottis. The vocal folds are placed between the thyroid cartilage and the arytenoid cartilages. All these elements, together with the top of the cricoid cartilage, comprise the larynx, where the sound is generated through the oscillatory movement of the vocal folds. Above the vocal folds, all the organs related to the speech production comprise the vocal tract, which consists mainly of the pharynx, the oral and nasal cavities, the tongue, the velum —or soft palate—, the hard palate, the teeth and the lips.

Speech sounds are rapid fluctuations in air pressure, generated when air is made to move by the vocal organs. These pressure fluctuations cause the listener's ear drum to move rapidly in an out. In this process, acoustic energy is transformed into mechanical energy at the eardrum. Then, this mechanical energy goes through several more transformations before arriving as patterns of neural energy at the listener's brain, where the information is processed resulting in the percept of sound.

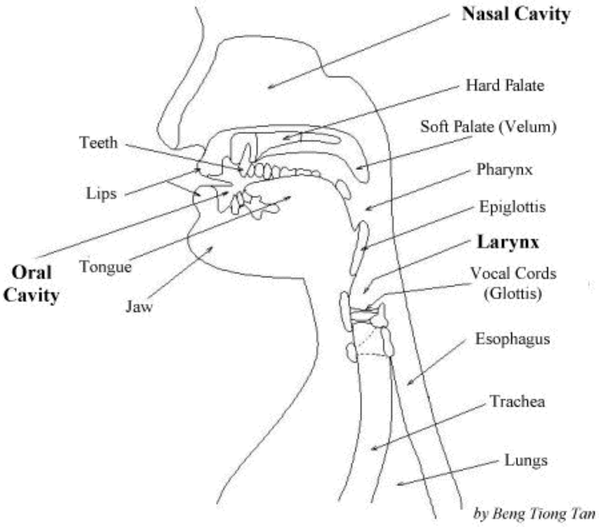


Figure 2.6: Human vocal system.

Physically speaking, the speech signal is a series of pressure changes in the medium. The oscillogram or waveform (air pressure changes in a speech wave as a function of time), the spectrum (amplitude plotted against frequency), and the spectrogram (amplitude, frequency and time plotted 3-dimensionally) are some of the most common representations of the speech signal. Figure 2.7 shows the waveform and the spectrogram corresponding to a sentence uttered by a female speaker; (as a reference, it is said that the fundamental frequency of general female voice is 225 Hz, male voice is 120 Hz and a small child's voice is 300 Hz (Kent and Read, 1992). Note that, in the spectrogram, amplitude is represented as a third dimension by dark shades.

The speech production system is usually modeled as the scheme in Figure 2.8, and described as two separate and independent processes: the sound generation in the larynx (*source*) on the one hand and the acoustic filtering of the speech sounds in the vocal tract (*filter*) on the other. This representation of the speech production is known as the source-filter approach. According to it, the voice can be modeled as an acoustic source signal in the larynx, filtered by a dynamic filter mimicking the supralaryngeal filter in the vocal tract, as mentioned above. Voiced sounds generated in the larynx are represented as a periodic pulse-train, while unvoiced sounds are

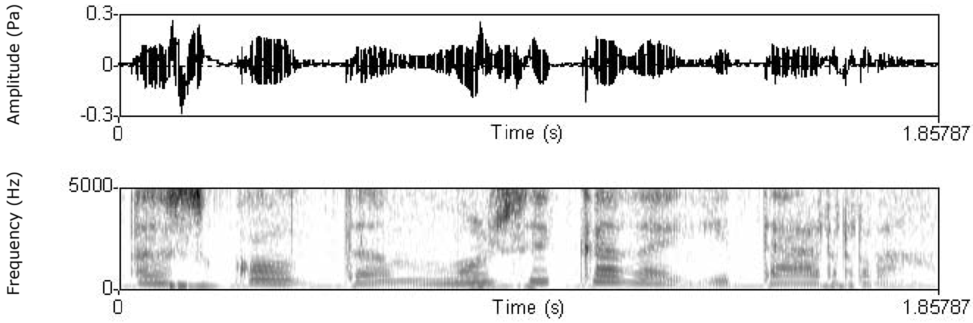


Figure 2.7: Waveform and spectrogram for a female voice speech sentence.

represented as noise (Fant, 1982).

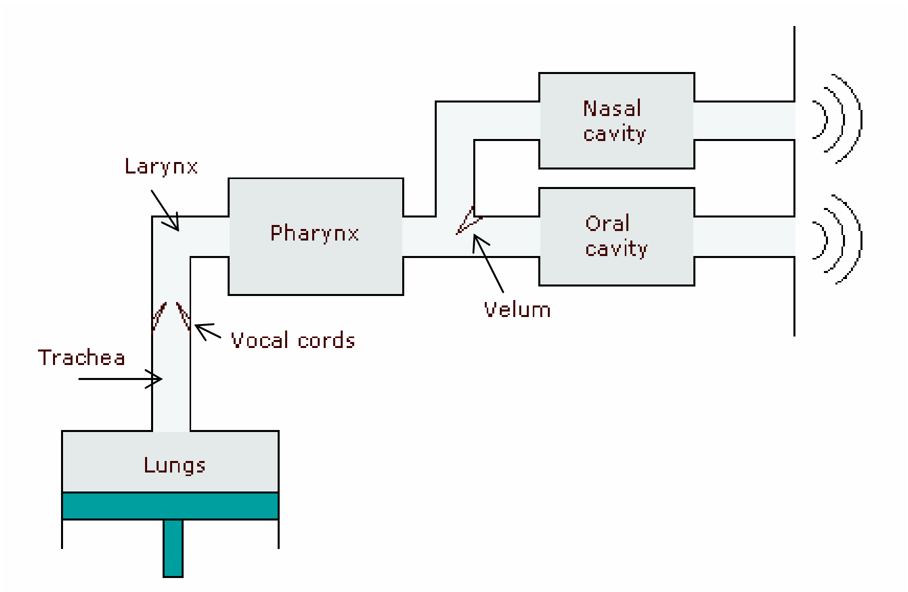


Figure 2.8: Schematic representation of the speech production system.

The linear model of speech production was developed by Fant in the late 1950s (Fant, 1960). The assumptions in this model are covered in detail by Flanagan (1972). The model assumes that, for a voiced speech *S(z)*, the source *U(z)* is a delta train, while for an unvoiced speech the source *U(z)* is represented by a random white noise. The filter is a cascade of a glottal pulse filter *G(z)*, a vocal tract filter *V(z)*, and a lip radiation filter *L(z)* (Figure 2.9).

Figure 2.10 shows two examples of spectra seen as a result of the laryngeal source combined with the vocal tract filter —or transfer— function. The top left diagram plots the spectrum of a glottal air flow with energy at the fundamental frequency (100 Hz) and the corresponding harmonics (200 Hz, 300 Hz, etc.). The bottom left diagram shows the corresponding spectrum for a fundamental frequency of 200 Hz. Note that the amplitude of the harmonics falls off gradually. The shape of the spectrum is mainly determined by the opening and closing movement of

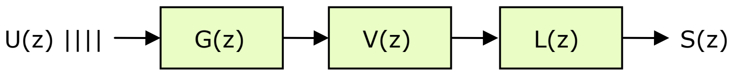


Figure 2.9: Linear speech production model.

the vocal folds. The supralaryngeal vocal tract performs as a time-varying acoustic filter that suppresses sound energy at certain frequencies while amplifying it at some other frequencies. Those frequencies at which there is maximum energy and the air vibrates with maximum amplitude are called *resonant frequencies* or *formants*, and they are partly determined by the overall shape, length and volume of the vocal tract. On the other hand, the shape of the filter function is determined by the whole vocal tract as an acoustic resonant system that includes losses due to radiation at the lips. The centre of the Figure 2.10 shows an idealised filter function with resonant frequencies at approximately 500 Hz, 1500 Hz and 2500 Hz. The diagrams placed at the right part of the figure show the spectra resulting from filtering the laryngeal source spectrum with the filter function. The laryngeal sources are clearly *shaped* by the filter functions.

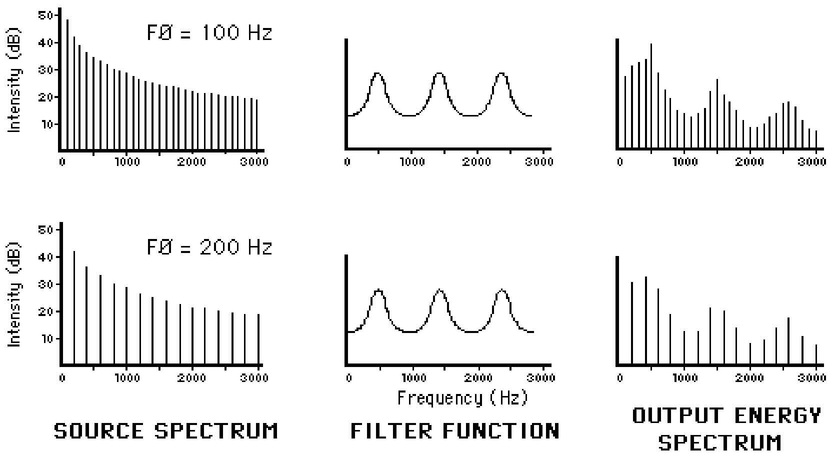


Figure 2.10: Output energy spectrum as a combination of a laryngeal source and a vocal tract filter function. (From http://www.haskins.yale.edu/featured/heads/MMSP/acoustic.html).

The physical size and shape of the speaker's vocal tract determine the range of sounds that can be produced by humans (Wolf, 1972; Hollien and Majewski, 1977; Rabiner and Schafer, 1978). Therefore, since each person has their own vocal characteristics and since most of them remain essentially unchanged even when the voice is disguised, there is reason to believe that an individual can be uniquely identified by voice alone (Wolf, 1972; Moosmuller, 1997).

Nevertheless, human voice is characterised by a high degree of variability within the same speaker —known as *intraspeaker variability*— so that two speech signals uttered by the same speaker are rarely equal, even when the speaker tries to make them identical. These differences can be caused by the speakers themselves due to different emotional states, colds, time of the day, age, etc., or by other external factors such as environmental noise, type of microphone, and channel distortion. Because of these changes in the speech signal, one of the goals of the speaker recognition technologies is to find those features that can properly characterise a single speaker; i.e. those features whose intraspeaker variability is smaller than their variability between speakers —also known as *interspeaker variability.*

Linguistic levels

Speech signal processing extracts features from the speech signal related to source and filter processes. Voice recognition systems tended to use only these filter parameters, which relate —in a complex way— to the physiology of the vocal tract and to the learnt articulatory configurations that shape the specific speech sounds (Rabiner and Juang, 1993; Gish and Schmidt, 1994; Campbell, 1997). These filter parameters are referred to as the spectral level of speech, since the frequency content (*spectrum*) of the acoustic wave is altered when the wave moves through the vocal tract.

Apart from the spectral content in speech, there are other acoustic parameters that have been shown to be useful in the speaker recognition task. Jitter and shimmer, for example, have been largely used to detect pathological and characteristic voices like breathy, rough or hoarse voices (Michaelis et al., 1998; Kreiman and Gerrat, 2005). More recently, they have also been used to determine the age and gender of the speakers (Wittig and Müller, 2005) and the classification of human speaking styles (Li et al., 2005).

Nevertheless, humans tend to use other information sources like lexical terms, prosody or phonetics to recognise others with voice. These levels of information are highly dependent on the learned habits or style, and they are mainly related to the use of linguistic cues derived from language. Therefore, they will be referred to as the linguistic level of speech.

At this point, the question might arise as to what a linguistic parameter really is. According to Rose (2002) and in terms of speech,

> a linguistic parameter can be thought of as any sound feature that has the potential to signal a contrast, either in the structure of a given language, or across languages and dialects.

Linguistic information can be divided into several levels or subfields, depending on the linguistic property analysed. The following table (2.3) relates each of these linguistic levels to their corresponding object of study.

Level	**Object of study**
Phonetics	Physical properties of sounds of human speech
Phonology	Sound system of a specific language or across languages
Morphology	Internal structure of words
Lexis	Words and phrases of a particular language
Syntax	Rules that govern the structure of grammatical sentences
Semantics	Meaning of words and phrases
Pragmatics	Use of utterance in communicative acts
Discourse analysis	Analysis of language in texts (spoken, written or signed)

Table 2.3: Linguistic levels and their corresponding object of study.

Speakers can also be characterised by their **dialect**, **sociolect** or **idiolect**. These are linguistic varieties which are differentiated from other varieties in terms of grammar, phonetics or lexis (Tusón, dir.). Dialect is related to a certain geographic region and sociolect is related to a certain social class, while idiolect refers to the use of language of a particular speaker. An idiolect includes dialectal and social characteristics, but also linguistic differences between individuals from the same region and social class. On the other hand, intonation, rhythm and stress are the three linguistic elements used to define and analyse **prosody** (Tusón, dir.).

Since speech carries some prosodic and other linguistic information that can be useful to recognise individuals, some speaker recognition systems have begun also to use the source parameters together with the filter parameters. These source parameters relate mainly to the fundamental frequency and power —or perceived pitch and loudness— of the speech and, in turn, to the prosody of the spoken phrases (Peskin et al., 2003; Reynolds et al., 2003). In section 2.3, a more detailed state of the art about prosody in automatic speaker recognition will be reviewed.

It has been seen that an individual's voice can provide information about the anatomy, physiology, age, linguistic aspects or even emotional states. The question arises whether the individual voice parameters can be easily mimicked by other speakers or not. Organic differences between speakers cannot be changed, so that when these differences are great, it may be difficult to achieve good imitations of another person's voice (Laver, 1994). However, many professional impersonators have shown to be successful in pretending to be someone else. A more detailed report about voice imitations and the effects of mimicking in automatic speaker recognition systems will be found in section 2.5.

2.2.3 Automatic speaker recognition techniques

In section 2.1.2, the general performance of a biometric recognition system was described. The current section is focused on the performance of speaker recognition systems as a particular case of a biometric system, emphasising the selection of appropriate features, along with the state-of-the-art techniques most commonly used to model the speakers, and the methods normally used in the classification and decision steps.

Feature extraction

Feature extraction or speech parameterisation consists in transforming the speech signal to a set of feature vectors (Bimbot et al., 2004). The aim of this transformation is to obtain a relatively low-dimensional representation, more suitable for statistical modeling and the computation of a distance or any other kind of score (in order to enable comparisons using simple measures of similarity) while preserving the information pertinent to the speaker.

Speaker recognition systems can be text-dependent or text-independent (Klevans and Rodman, 1997). In a text-dependent application, each speaker reads a short prescribed text to train the system. Then, during the testing phase, the unknown speakers have to pronounce the same text that was used in the training phase. On the other hand, text-independent applications allow the speaker to use any text in both training and testing phases. This approach is based on the hypothesis that acoustic parameter measurements of individual speakers uttering any speech may be used to characterise the speaker uniquely (Higgins et al., 1996; Rodman, 1998).

Nevertheless, voice is subject to variations caused by several factors: speaker's physical characteristics, health conditions, emotional states, dialect, sociolect and idiolect spoken by the individual, age, or even the recording and transmission conditions. In Lee et al. (1999), for example, the authors analyse the changes in magnitude and variability of duration, fundamental frequency, formant frequencies and spectral envelope of children's speech as a function of age and gender.

In order to guarantee a correct performance of the system, the parameters used to describe the speaker's voice should be characterised by the following desired conditions (Wolf, 1972):

- low variability within the same speaker
- good discrimination between speakers
- low variability over time
- independent on health conditions
- not subject to mimicry
- unaffected by environmental and transmission noise
- easily measurable
- occur naturally and frequently in speech

In order to find good features for speaker recognition, a method known as ANOVA —which comes from ANalysis Of Variance— can be used (Campbell, 1997). This technique measures the Fisher's F-ratio (Equation 2.1) between the sample pdf's of different features:

$$F = \frac{\text{variance of speaker means}}{\text{average intraspeaker variance}} \tag{2.1}$$

High F-ratios are obviously desirable for speaker recognition, but F-ratio has to be evaluated for many different combinations of features to be useful; two correlated features with high individual F-ratios might be less effective as a feature vector than two uncorrelated features with low F-ratios.

It was seen in section 2.2.2 that the spectral shape of the voice signal contains information about the vocal tract and the excitation source in the glottis by means of the formants and the fundamental frequency, respectively. Therefore, most of the parameters used are obtained from the spectrum of the signal. The spectral parameters obtained for the speaker recognition analysis are usually the same as the ones used in speech recognition techniques, although the objective of both applications is not the same.

The process of extracting these spectral parameters from the voice signal is divided in different small and consecutive processes. First of all, the voice signal is converted into an electrical signal by a microphone. The obtained electrical signal is then sampled and quantised in order to obtain a new digital signal. Next, the signal is segmented in temporal frames, whose length is typically 20–30 ms with 10 ms time shift. The frames are weighted by a window (usually a Hamming window). Finally, a spectral estimation is done in each segment, and the parameters are extracted (Figure 2.11).

The most commonly used parameters in the state-of-the-art speaker and speech recognition technologies are the Mel-frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980; Oppenheim, 2004), obtained as shown in Figure 2.12. First of all —although optionally— a first

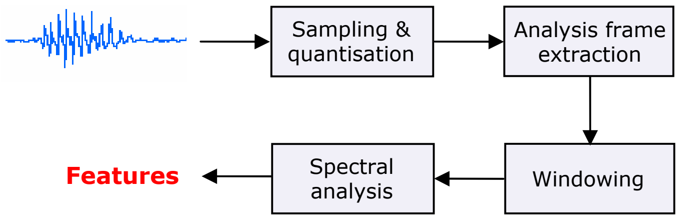


Figure 2.11: Spectral coefficients extraction process.

order high-pass filter is applied in order to emphasise the high frequencies of the signal. Next, the signal is segmented in temporal frames, using the common values mentioned above; then, applying a window —typically a Hamming window— in order to reduce the discontinuities of the segmentation, a Discrete Fourier Transform (DFT) is performed. The resulting values are then passed throug a filterbank distributed along the frequency domain according to a *mel scale*, obtaining a vector of energy values: the Filter Bank Energies (FBE).

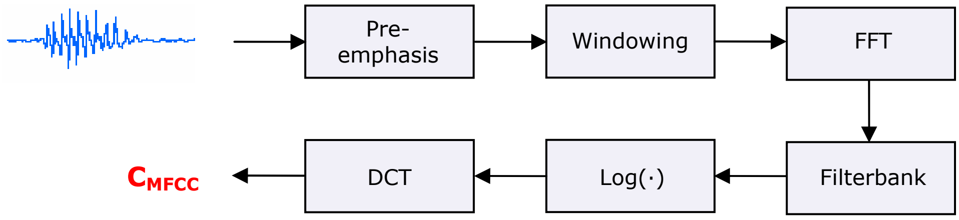


Figure 2.12: Mel-frequency cepstral coefficients extraction process.

The mel scale, proposed by Stevens (1937), is based on the manner how the speech perception works in the human ear. The human auditory system resolves frequencies non-linearly across the audio spectrum; hence, empirical evidence suggests that a system that operates in a similar non-linear way, obtaining the desired non-linear frequency resolution, provides a better recognition performance.

The mel scale filterbank is a series of Q triangular bandpass filters that have been designed to simulate the bandpass filtering according to the the human auditory system. The series of triangular filters are 50% overlapped and with constant bandwidth and spacing on a mel frequency scale. On a linear frequency scale, this filter spacing is approximately linear in the range 0 to 1000 Hz and logarithmic at higher values of frequency. The triangles are all normalised so that they have unit area.

A filterbank representation is illustrated in Figure 2.13, where Q is the number of filters (usually between 20 and 24). By applying this filterbank, the number of coefficients is reduced, so that the information is compacted, and variance is also reduced when averaging the samples of the DFT in each filter. Finally, a logarithmic compression and the Discrete Cosine Transform (DCT) is applied to the vector of FBE in order to obtain the MFCC:

$$C_{MFCC}(i) = \sum_{k=1}^{Q} \log\left(E(k)\right) \cos\left(\frac{i(k-0.5)\pi}{Q}\right), \qquad 0 \le i < N_c \tag{2.2}$$

where i is the index of the MFCC, k is the index of the filter, $E(k)$ is the energy of the k filter, and N_c is the total number of coefficients.

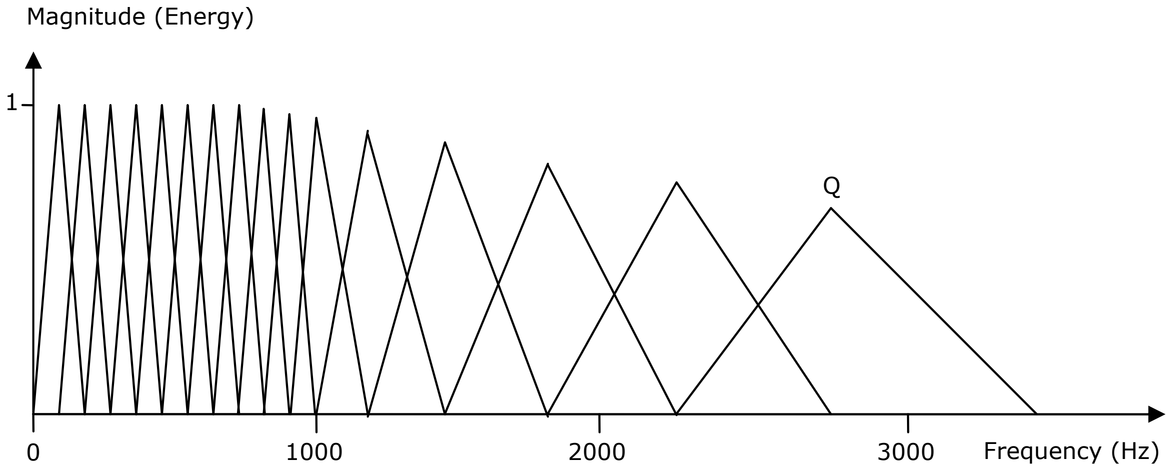


Figure 2.13: Example of a mel-scale filterbank representation for a set of 16 coefficients.

MFCC are the parameters of choice for many speaker and speech recognition applications. Nevertheless, other cepstral parameters can be obtained by a codification process by means of a linear prediction (Picone, 1993). These parameters are known as Linear Prediction Cepstral Coefficients (LPCC), but in terms of characteristics, they do no differ considerably from the MFCC ones.

On the other hand, in Nadeu et al. (1995) the Frequency Filtering (FF) method is presented. In this technique, a first order filter $H(z) = z - z^{-1}$ is used instead of the Discrete Cosine Transform in the MFCC extraction process. The use of this filter consists in a simple subtraction of the energies of two bands to compute each parameter, so that the computational cost is lower and the resulting parameters remain in the frequency domain. In most of the experiments performed in Hernando (1997) and Abad et al. (2003), these parameters give comparable or even better results than mel-cepstrum coefficients.

The above described parameters are known as static coefficients since they do not take into account the temporal variation along the frames. On the other hand, it has been commonly assumed that the feature vectors of consecutive frames are uncorrelated; however, this is not true, since the spectrum of the signal in a frame depends, in some way, on the spectrum in its consecutive frames. Moreover, the spectrum variation in the time domain is able to provide some useful information about the speaker identity. Therefore, this information has been widely added in the state-of-the-art systems (Furui, 1986). These additional parameters, known as *dynamic coefficients* or *deltas*, are computed by the following linear regression:

$$\Delta c(l,k) = \frac{\sum_{d=1}^{D} d(c(l+d,k) - c(l-d,k)}{2\sum_{d=1}^{D} d^2} \tag{2.3}$$

where $c(l,k)$ is the static coefficient k in the frame l, and D is the window size used to compute the dynamic parameters. Applying Equation (2.3) again over these delta parameters, their temporal variation is obtained in terms of new dynamic parameters known as *accelera-*

tion or *delta-delta* coefficients. Both dynamic parameters —delta and delta-delta— are usually concatenated to the static coefficients obtaining a single feature vector.

Automatic speaker recognition systems have relied mostly on the short-term features described above. However, there are other levels of information that convey useful information about the speaker. Recent studies (Carey et al., 1996; Sönmez et al., 1998; Doddington, 2001; Andrews et al., 2002; Bartkova et al., 2002; Weber et al., 2002; Peskin et al., 2003; Reynolds et al., 2003) have demonstrated that these levels can add complementary knowledge to the traditional spectrum-based recognition systems, improving their accuracy.

Research on additional information sources in speaker recognition has been mainly focused on the use of the fundamental frequency. One of the reasons is that, like other linguistic-related parameters, it appears to be more robust to acoustic degradations from channel and noise effects (Atal, 1972; Carey et al., 1996). Arcienega and Drygaljo (2001), for example, suggest the use of F0-dependent speaker models. In Sönmez et al. (1998) and Adami and Hermansky (2003), the variation of fundamental frequency over time is modeled for use in a speaker recognition task, together with the signal energy variation (in Adami and Hermansky (2003)).

Apart from extracting parameters from the fundamental frequency and energy, other studies suggest the use of different information sources: number of phonemes per word, number of frames per word, pause rate and duration, etc. (Peskin et al., 2003). In Klusácec et al. (2003), the phonemes output by a speech recognition system using acoustic models of entire words are compared with the phonemes output by a recognition system using acoustic models of phonemes, in order to know how the speaker pronounces the words. In Jin et al. (2003), the voice signal is processed with several speech recognition systems in different languages, and the outputs are compared in order to recognise the speakers.

Several modeling techniques that use this additional information sources will be reviewed at the end of this section. Furthermore, next section (2.3) will deal with the use of prosody in speaker recognition, focusing on the extraction of prosodic parameters and the state-of-the-art of prosodic modeling techniques.

Speaker modeling and pattern matching

In order to enrol users into a recognition system, models corresponding to each speaker (user) are built from the set of features extracted from the speech signal and stored in database. The pattern matching task consists in calculating a match score, which is a measure of the similarity between the input feature vectors and a stored model. In the user authentication step, the score corresponding to the input speech signals is compared to the model of the claimed user.

When performing the pattern matching task, two main types of models can be found: template models and statistical —or stochastic— models. The pattern matching for template models is deterministic: the observation (a feature vector of a collection of vectors from the unknown speaker) is seen as an imperfect replica of the template, and the alignment of observed frames to template frames is selected to mimimise a distance measure d. On the other hand, pattern matching in statistical models is probabilistic and results in a measure of the likelihood —or conditional probability— of the observation given the speaker model (Campbell, 1997).

The match score between the template $\bar{x}$ for the claimed speaker and an input feature vector x_i from the unknown speaker is given by a distance $d(x_i, \bar{x})$. Many different distance measures between these two vectors can be expressed as:

$$d(x_i, \bar{x}) = (x_i - \bar{x})^T W (x_i - \bar{x}) \tag{2.4}$$

where W is a weighting matrix. If W is an identity matrix, for example, the distance is *Euclidean*; if W is the inverse covariance matrix corresponding to mean $\bar{x}$, the distance is then known as the *Mahalanobis* distance.

A more detailed description about general distances can be found in Basseville (1989). Next, some of the state-of-the-art pattern matching methods are introduced. Techniques like dynamic time warping and vector quantisation are based on template models, while statistical models can be found, for example, in hidden Markov models.

Long-term averaging Long-term averaging of features extracted from the speech signal —in both time and frequency domains— was one of the first modeling techniques for speaker recognition. The method consists in obtaining a large number of feature vectors from each known speaker, and compute the mean and variance of each component of the feature vector for all the samples of a speaker. Then, the similarity between speakers is determined by calculating a weighted distance measure between the average feature vectors of two speakers (Markel and Davis, 1978).

Intraspeaker variability increases a lot in short sentences; so, when using long-term averaging approach, the accuracy of the systems is highly dependent on the amount of training and testing data (Wrench, 1981). Nevertheless, long-term averaging has been used with several kinds of features: pitch, cepstral coefficients, inverse filter spectral coefficients, etc. (Shridhar et al., 1981; Doddington, 1985; Campbell, 1997).

Dynamic time warping Dynamic time warping (DTW) is the most popular method to compensate for speaking-rate variability in template-based systems (Sakoe and Chiba, 1978). A text-dependent template model consists of a sequence of templates $(\bar{x}_1, ..., \bar{x}_N)$ that must be matched to an input sequence $(x_1, ..., x_M)$, where N is normally not equal to M because of timing inconsistencies in speech. A DTW algorithm does a constrained, piece-wise linear mapping of one or both time axes to align the two speech signals while minimising the accumulated distance z, which is expressed as follows:

$$z_{DTW} = \sum_{i=1}^{M} d(x_i, \bar{x}_{j(i)}). \tag{2.5}$$

The template indexes $j(i)$ are given by the DTW algorithm. At the end of the time warping, this accumulated distance z_{DTW} is used as a match score.

Dynamic time warping accounts for the variation over time of parameters corresponding to the dynamic configuration of the articulators and vocal tract (Campbell, 1997).

Vector quantisation Vector quantisation (VQ) uses multiple templates to represent frames of speech; rather than a single cluster of data for each speaker's model, this technique segregates data into multiple clusters and determines their centroids. A VQ codebook is created by standard clustering procedures for each known speaker using his training data. In the testing phase, the pattern match score is the distance between an input vector and the minimum distance code word in the VQ codebook C. According to this, the matching score for L frames of speech is given by the following expression:

$$z_{VQ} = \sum_{j=1}^{L} \min_{\bar{x} \in \mathrm{C}} (d(x_j, \bar{x})) \tag{2.6}$$

The clustering procedure used to design the codebook averages out temporal information from the code words; therefore, a time alignment is no needed. As a consequence, speaker-dependent temporal information, which may be present in the spoken sentences, is neglected (Linde et al., 1980; Soong et al., 1985; Campbell, 1997).

Hidden Markov models A hidden Markov model (HMM) is a stochastic model commonly used for modeling sequences, in which the observations are a probabilistic function of the state (Rabiner, 1989; Rabiner and Juang, 1993; Campbell, 1997). The HMM is a finite-state machine, where a *pdf* (or feature vector stochastic model) $p(x|s_i)$ is associated with each state s_i (the hidden underlying model). The states are connected by a transition network, where the state transition probabilities are $a_{ij} = p(s_i|s_j)$. Figure 2.14 illustrates an example of a three-state hidden Markov model:

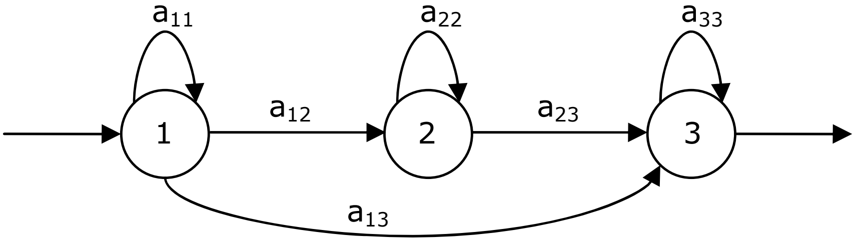


Figure 2.14: Example of a three-state hidden Markov model.

By using the Baum-Welch algorithm, the probability that a sequence of speech frames was generated by this model can be determined (Rabiner and Juang, 1986; Rabiner, 1989). This probability —or likelihood— is used as a score for L frames of input speech given the model (Rabiner and Juang, 1986; Rabiner, 1989; Campbell, 1997):

$$p(x(1;L)|model) = \sum_{\substack{all\,state \\ sequences}} \prod_{i=1}^{L} p(x_i|s_i) p(s_i|s_{i-1}) \tag{2.7}$$

Gaussian mixture models A Gaussian mixture model (GMM) is a weighted sum of Gaussian density functions that models the distribution of the feature vectors extracted from the speech signal (Reynolds, 1995; Reynolds and Rose, 1995). Given a D-dimensional feature vector $\mathbf{x}$, the Gaussian mixture model λ_i corresponding to the speaker S_i is defined by the following expression:

$$p(\mathbf{x}|\lambda_i) = \sum_{m=1}^{M} \omega_m N(\mathbf{x}, \mu_{im}, \Sigma_{im}) \tag{2.8}$$

where M is the number of components, w_m is the weight corresponding to the component m, and $N(\mathbf{x}, \mu_{im}, \Sigma_{im})$ is a Gaussian function defined as follows:

$$N(\mathbf{x}, \mu_{im}, \Sigma_{im}) = \frac{1}{(2\pi)^{\frac{D}{2}}\sqrt{|\Sigma|}} \exp(-\frac{1}{2}(\mathbf{x} - \mu_{im})^T \Sigma_{im}^{-1}(\mathbf{x} - \mu_{im})) \tag{2.9}$$

where μ is the vector of means and Σ the covariance matrix. A GMM is, in fact, a one-state HMM.

In the recognition step, given a sequence of test feature vectors $X = [x_1, x_2, ..., x_T]$ extracted from an unknown user's speech, the probability of the unknown speaker being the speaker S_i (assuming that vectors x_i are independent) is determined by the following expression:

$$p(\mathbf{X}|\lambda_i) = \prod_{t=1}^{T} P(x_t|\lambda_i) \tag{2.10}$$

which will be used as a matching score.

Artificial neural networks Artificial neural networks (ANN) are also used in speaker recognition applications. The kind of neural networks used are feed-forward neural networks, where the information moves only in one direction (forward) from the input nodes, through the hidden nodes, if any, and to the output nodes. A feed-forward neural network is created for each known speaker, and each network contains one output that is trained to be active only for its speaker. In the testing phase, an input feature vector is fed forward through each network, and the identification is determined by the network with the highest accumulated output values. In the speaker verification mode, the input vectors of the unknown user are fed forward through the network belonging to the claimed speaker. If the average output value is bigger than a threshold, the speaker is accepted (Oglesby and Mason, 1990).

Rudasi and Zahorian (1991) demonstrated that by using small binary networks for distinguishing between two speakers instead of one large network with one output for each known speaker, the performance in speaker recognition was much better, since the binary networks were much more specialised. Another kind of networks, the time-delay neural networks (TDNN), were developed by Bennani and Gallinari (1991) to capture transient information using a connectionist approach.

The decision step

The previous sections have dealt with the most commonly used methods to extract relevant features from the speech signal and the way how the corresponding speaker models are formed. Once the scores or probabilities have been obtained, these are used to take a recognition decision, which represents the final phase of a speaker recognition process.

As it was seen above, models can be classified into two types: the template models and the statistical —or stochastic— models. When using template models, the likelihood can be approximated by exponentiating the match scores, which are given by the distance measure d (Campbell, 1997):

$$L = \exp(-ad) \tag{2.11}$$

where a is a positive constant. This is equivalent to assume that scores are proportional to log likelihoods.

The **nearest neighbour** (NN) algorithm has been widely used in identification applications based on template models (Higgins et al., 1993; Jaeger et al., 2005). Given a test utterance, the aim of this classifier is, first, to determine the closest model or template —i.e. the *nearest neighbour*— and second, to determine which is the corresponding model —in the identification case— or if the utterance corresponds or not to a claimed model —in the verification case— by using that closest neighbour (Cunningham and Delany, 2007). Since more than one neighbour is normally taken into account, the technique is commonly referred to as k-nearest neighbour (kNN), where k nearest neighbours are used in the identification process. Majority voting and sum rules are the most commonly used approaches in NN classification (Campbell, 1997; Kou and Gardarin, 2002; Cunningham and Delany, 2007).

In statistical models like GMM, for example, likelihood measure is given by the probability of the unknown speaker having generated the observation (utterance) $\mathbf{X}$:

$$L = p(\lambda_i|\mathbf{X}) \tag{2.12}$$

In speaker identification applications, given a set of S speakers $\{s_1, s_2, ..., s_S\}$, and a set of models for each speaker $\{\lambda_1, \lambda_2, ..., \lambda_S\}$, the aim is to find the speaker whose model is assigned to the maximum likelihood:

$$\hat{L} = \arg \max_{1 \leq i \leq S} L \tag{2.13}$$

Bayes rule can be applied to statistical models combining Equations 2.12 and 2.13 obtaining the following expression:

$$\hat{L} = \arg \max_{1 \leq i \leq S} \frac{P(\mathbf{X}|\lambda_i)P(\lambda_i)}{P(\mathbf{X})} \tag{2.14}$$

Since $P(\mathbf{X})$ remains constant for all the speakers, the maximisation is not affected by this probability. Moreover, all the speakers are assumed to be equally possible, so that the decision rule can be simplified as follows:

$$\hat{L} = \arg \max_{1 \leq i \leq S} P(\mathbf{X}|\lambda_i) \tag{2.15}$$

where the probability $P(\mathbf{X}|\lambda_i)$, when using Gaussian mixture models, is given by Equation 2.10.

In speaker verification applications, the aim is to determine if an unknown speaker corresponds or not to a claimed speaker. The test speech signal of the unknown user is compared with the model of the claimed speaker, obtaining a similarity score as a result: the higher the score, the more similar the user and the claimed speaker. Let then an unknown user claim to be the speaker S_i with a model λ_i, generating an observation $\mathbf{X}$. A decision can be taken by choosing a threshold θ_i. If the score value is higher than the threshold, the user will be accepted; on the other hand, if the score lies below the threshold, the user will be rejected:

$$p(\mathbf{X}|\lambda_i) = \begin{cases} \geq \theta_i & \text{accepted;} \\ < \theta_i & \text{rejected.} \end{cases} \tag{2.16}$$

The threshold θ_i can be speaker-dependent or independent. However, it is rather difficult to set a fixed value of θ_i, since the scores of the same speaker present some differences due to the intraspeaker variability. In order to reduce such variability, scores are usually normalised before taking the final decision.

Given a match score, the speaker verification problem involves choosing between two hypothesis:

- H_1: the user is the claimed speaker S_i (the user is a *client*); hence, the speech signal $\mathbf{X}$ was generated by S_i.
- H_0: the user is not the claimed speaker S_i (the user is an *impostor*); hence, the speech signal $\mathbf{X}$ was not generated by S_i.

With the probability functions of both hypotheses assumed to be known, the optimal solution is given by the *likelihood ratio* (Reynolds et al., 2000):

$$LR(\mathbf{X}|S_i) = \frac{P(\mathbf{X}|H_1)}{P(\mathbf{X}|H_0)} \begin{cases} \geq \theta_i & \text{choose } H_1; \\ < \theta_i & \text{choose } H_0. \end{cases} \tag{2.17}$$

where $P(\mathbf{X}|H_i)$ is the probability of $\mathbf{X}$ given the hypothesis H_i. As in Equation 2.16, if the likelihood ratio (LR) lies above a specific threshold, the hypothesis H_1 will be accepted —and viceversa.

In order to compute $LR(\mathbf{X}|S_i)$, both λ_i and λ_{imp} models are needed to represent H_1 and H_0 hypotheses, respectively. Since it is rather common, in this case, to work in the logarithmic domain, LR results in the following *log-likelihood ratio* expression:

$$LLR(\mathbf{X}|S_i) = \log(P(\mathbf{X}|\lambda_i)) - \log(P(\mathbf{X}|\lambda_{imp})) \begin{cases} \geq \theta_i & \text{choose } H_1; \\ < \theta_i & \text{choose } H_0. \end{cases} \tag{2.18}$$

where the second term of the Equation is a kind of normalisation, which provides a more stable threshold (Matsui and Furui, 1995).

The model λ_{imp} must contain every possible alternative to the speaker S_i. Such an impostor model can be built as following two different approaches. The first approach consists in using a set of several speaker models —or *cohort speakers*— different to the speaker S_i (Reynolds and Rose, 1995; Rosenberg and Parthasarathy, 1996). In order to improve the system performance, a set of cohort speakers can be assigned to each trained speaker.

On the other hand, a second approach consists in training the impostor model as a single model using several speakers (Matsui and Furui, 1995; Rosenberg and Parthasarathy, 1996; Reynolds, 1997). This model is usually called *universal background model* (UBM), and when using Gaussian mixture models, the method is known as the GMM-UBM approach. When template models are used instead of stochastic models, likelihood ratios can also be formed using global speaker models to normalise L(Equation 2.11).

Additional modeling techniques using linguistic information

It is well-known that speakers normally use distinctive vocabulary in their individual speech, and such particular words or phrases can be used by human listeners to identify individuals. On the basis of this assumption, George Doddington tried to identify speakers by using only conversation transcriptions alone, with no reference to the sound of the speakers' voice. Doddington made a breakthrough showing that long-term speech characteristics, particularly the frequent usage of certain words or phrases, contained useful information for characterising speakers (Doddington, 2000, 2001). In this preliminary work, statistical language models were modeled using simple word-level unigram and bigram frequencies.

These particular characteristics —named *idiolectal* differences— motivated further similar experiments performed by García-Romero et al. (2003), in which some idiolectal information was provided by using lexical bigrams in the language models in order to improve the performance of an acoustic based GMM system.

One of the most relevant problems when using prosodic features in speaker recognition systems is the large amount of data required for training. Prosodic features are sparser than short-term frame-level features, so that they may be under-trained without a considerable amount of data. Therefore, after Doddington showed that idiolect provided competitive speaker recognition performance given a sufficient amount of data, NIST introduced the Extended Data Task (Doddington et al., 2000) for the 2001 NIST Evaluation.

The NIST Extended Data Task contained the complete Switchboard-I corpus (Godfrey et al., 1990) of conversational telephone speech for training and testing material. A total of 2430 conversations were recorded by 543 English native speakers, and each conversation involved a different speaker pairing. Along with the audio data, NIST provided transcripts for the complete corpus.

Motivated by the success of Doddington's preliminary work, the *SuperSID* project was undertaken at the 2002 Summer Workshop on Language Engineering at the Johns Hopkins University (Reynolds et al., 2002) in order to improve the accuracy of automatic speaker recognition technology by exploiting new information sources in conversational speech. SuperSID experiments were focused on extracting and modeling speaker-specific patterns in acoustic cues, prosody, word and phone pronunciations, word usage and interactions with conversation partners (Adami and Hermansky, 2003; Adami et al., 2003; Campbell et al., 2003; Jin et al., 2003; Klusácec et al., 2003; Navrátil et al., 2003; Peskin et al., 2003; Reynolds et al., 2003).

Following the work of Doddington, in which idiolectal differences among speaker were modeled by means of n-grams, Andrews et al. (2001) performed a set of experiments based on n-gram modeling of phonetic units. The results showed that the phonetic models provided complementary speaker information to the short-time acoustic based system. Word usage models were also used in SuperSID experiments in order to assess the impact of speech recognition errors on speaker identification performance. By using four versions of transcriptions corresponding to the Switchboard database, it was found —as expected— that the speaker recognition performance degraded as speech recognition diminished. Nevertheless, word-based speaker recognition seemed relatively robust to automatic speech recognition error, and speaker recognition systems based on lexical features provided one of the strongest sources of complementary information.

Regarding conversational style, features based on speaker turns in a dialogue were analysed in order to see whether the speaker interaction style observed in a conversation contained useful complementary information to the speaker recognition task. Furthermore, a set of experiments were performed in order to see whether turn-based features and the lexical information from the non-target side of the conversation could also help to improve the performance in speaker recognition. The distribution of words in the target side was modeled using the n-gram model, and then conditioned on a function of the word distribution in the non-target side. Turn-based features —consisting of phone rate, average low energy and average $\log F0$— from the target side of the conversation were found to improve the speaker recognition system, whereas the non-target side features were not helpful to improve performance, either for the turn-based or the word-frequency experiments.

By using all these additional linguistic information sources —acoustic cues, word and phone pronunciation, word usage and interactions with conversation partners— together with a set of prosodic features, more than a seventy percent relative improvement —in terms of equal error rate reduction— was achieved over existing approaches on a standard NIST Speaker Recognition Evaluation Task.

2.3 Prosody in speaker recognition

As it was seen in section 2.2.2, humans use several information levels to recognise others with voice. The voice timbre, a characteristic laugh or a repeatedly used term, for example, are some of the factors that can help in the human recognition process (Campbell et al., 2003; Reynolds et al., 2003). These speaker-specific characteristics contained in speech are mainly related to physical and learned sources. Depending on the speech source, several information levels can be considered, and their related features are extracted in order to use them in the speaker recognition process.

This section focuses on the use of the linguistic level —and, in turn, on the use of prosody— in the speaker recognition process. First, the prosodic elements that can be found in human speech are briefly described. Second, the most commonly used prosodic models in the speaker recognition task are briefly reviewed.

2.3.1 Prosodic elements in speech

Prosody is conveyed through three different elements: intonation, rhythm and stress (Tusón, dir.), which are perceived by listeners as changes in fundamental frequency, loudness and sound duration, respectively (Adami, 2007). Humans use different methods to produce these linguistic phenomena that determine prosody (Ladefoged, 1968; Collier, 1975). Changes in the respiratory system and laryngeal muscles, for example, have an important role in controlling the fundamental frequency (Lehiste, 1970; Atkinson, 1978; Boves and Strik, 1988).

Fundamental frequency is determined physiologically by the number of cycles that the vocal folds make in a second, and they are the natural result of the length of these structures. On the other hand, vocal intensity is directly related to the subglottic pressure of the air column. This subglottic pressure, in turn, depends on several factors such as amplitude of vibration and tension of vocal folds (Wertzner et al., 2005). Both prosodic elements are somehow related; variations of intensity, for example, depend on frequency (Coleman et al., 1977), since the increasing in laryngeal tonus generates a higher glottic pressure and, consequently, more intensity. For this reason, high voices tend to be more intense (Wertzner et al., 2005).

It is well-known that prosody plays and important role in the speech act and communication in everyday speech (Nooteboom, 1997; Wennerstrom, 2001). The speech, in turn, becomes adjustable to the the particularities of the speaker, so that each person may use distinct variations of tone, intensity and rhythm in their speech production. This changes, manifested in the speech prosody, are essential to give melody to the spoken sentences (Wertzner et al., 2005).

Although prosody is mainly determined by the variations of fundamental frequency, intensity and sound duration, the precise definition of prosody can differ between languages. Many languages —like Catalan, Italian and English, for instance— use the fundamental frequency to express emotions like surprise, sadness, irony, etc., or to distinguish statements from questions. These languages are called *intonation (or intonational) languages*. Figure 2.15 shows an example of a spoken sentence in an intonation language (English), where fundamental frequency and intensity contours are depicted in blue and green colours, respectively. The utterance corresponds to a female voice pronouncing the question: *are you ready now?*

On the other hand, some languages use fundamental frequency semantically to distinguish words. Such languages —like, for instance, Chinese, Somali, Thai or Hausa— are known as *tonal (or tone) languages*. Other languages use a tonal word accent, so that, in terms of tonality, they occupy an intermediate position between intonational and tonal languages. Examples of these languages are Norwegian, Swedish and Japanese. Pitch is also changed in some African languages for grammatical purposes. In Luo language, for example, change in fundamental frequency is used to mark the past tense. However, tonality is not the only linguistic characteristic used in a semantic way to distinguish the meaning of words. Some languages, for example, use vowel duration to make lexical distinctions.

So, the use of prosodic elements can vary a lot across languages. In intonational languages, syllables are given an *accent*, which consists of a louder intensity or *stress*, a lengthened word and a higher fundamental frequency, or some combination of these prosodic characteristics. On the other hand, in tonal languages the pitch accent must be present, but the other prosodic parameters are optional.

Variations in sound duration, fundamental frequency and stress —or intensity— normally apply to more than one phoneme: syllables, words, phrases, clauses, etc. Since phonemes are known as speech segments in linguistic terms, these prosodic elements (rhythm, stress and F0) are also know as *suprasegmental features*, and they are usually analysed over sequences of segments or entire syllables (Tusón, dir.; Dellwo et al., 2007).

These suprasegmental features go beyond other linguistic levels like lexis or semantics, and they can be perceived without knowing the lexical or semantic content of the spoken message. Prosody is used, for example, when listening to a conversation through a wall, or when trying to discern the target speaker of comedians' imitations (Kajarekar et al., 2003).

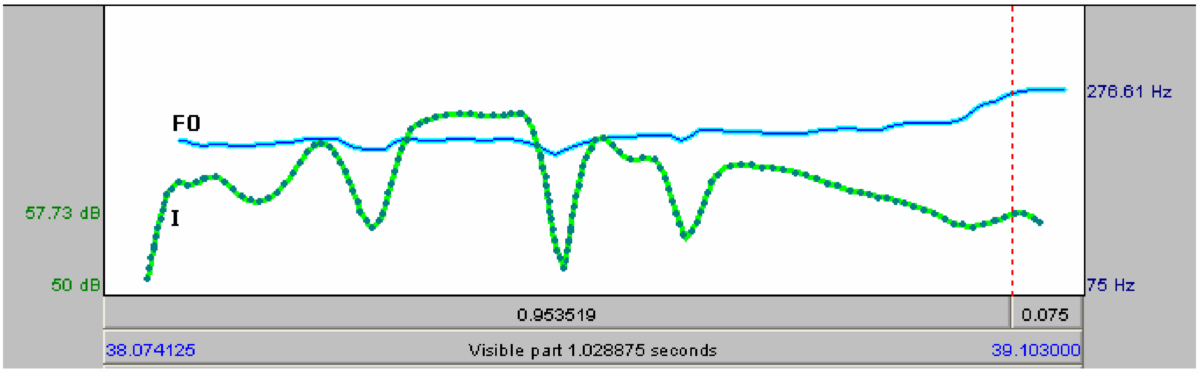


Figure 2.15: Fundamental frequency (in blue) and intensity (in dotted green) contours of a female utterance: *are you ready now?*.

Prosody is also used to capture different expressions of emotional states. Nevertheless, since speech is a cultural activity, signs of emotion in speech may also be subject to cultural influences (Cowie et al., 2001; Douglas-Cowie et al., 2003). Joy, elation, and hot anger, for example, seem to be rather universal, while other states like sadness, fear and boredom have usual inconsistences in F0 and speech rate features.

2.3.2 Prosodic models

Since prosody plays an important role in the human recognition process, the use of prosodic features has been investigated during the last decades, and a lot of effort has been placed in using this kind of information to automatic speaker recognition systems.

Normally, in most of the experiments performed, systems based only on prosodic parameters do not outperform the traditional filter-based systems. Nevertheless, several studies have demonstrated that prosodic features can be used to effectively improve the performance of the conventional systems based solely on filter parameters, supplying complementary information not captured in the traditional systems (Carey et al., 1996; Sönmez et al., 1998; Doddington, 2001; Andrews et al., 2002; Bartkova et al., 2002; Weber et al., 2002; Peskin et al., 2003; Reynolds et al., 2003). Therefore, these features have mainly been used as complementary parameters to the state-of-the-art spectral systems.

Moreover, some of these features have the advantage of being more robust to noise than the spectral ones (Carey et al., 1996). Spectral patterns can be affected by frequency features of the transmission channel, and spectral information depends also on the speech level and the distance between the speaker and the array, while fundamental frequency is unaffected by such variations (Atal, 1972).

According to Adami (2007), prosodic features can be modeled in three different ways: by *overall distribution statistics*, by *contour statistics* and by *contour matching*. Next, these three classes of modeling are reviewed.

Overall distribution statistics

Systems based on this approach use the distribution statistics of prosodic features estimated over all the training data. The most common example is comparing the mean and standard deviation of the fundamental frequency between enrolment and test utterances; in this way, Carey et al. (1996) demonstrated that the mean and variance of fundamental frequency and energy variance provide discriminatory information about the speaker when used individually. Using also an overall distribution statistics approach, Sönmez et al. (1997) proposed a probabilistic model of pitch halving/doubling in order to model speaker information.

The prosodic feature obtained by means of this approach can be appended to a vector of standard spectral based features and used in traditional modeling systems. The most relevant problem by using this statistical approach is that the temporal dynamic information of the prosodic features is not adequately captured. This lack of information has been overcome by using statistics of feature time derivative and dynamic features derived from segments.

Contour statistics

Contour statistics speaker modeling is based on statistical measurements estimated from segments of speech. Atal (1972) and Markel et al. (1977), for example, extracted the average of pitch and intensity of fixed-size segments to model speaker information. Atal applied dynamic time warping to compare F0 contours between two utterances of the same text, so that it was able to capture speaker-specific temporal dynamic events. Other approaches have been proposed using variable-length size segments to estimate the statistics (Sönmez et al., 1998; Kajarekar et al., 2003), where a F0 stylisation algorithm and the speech pauses are used to detect the segments, and Shriberg et al. (2005) computed several duration, fundamental frequency and energy features for each estimated syllable in speech recognition output.

Nevertheless, the estimated statistics from these segments do not capture differences in the realisation of prosodic features in an adequate way; for example, different contour shapes within a segment can yield the same statistics. In order to overcome it, Adami et al. (2003) proposed a quantisation of the slopes of the energy and fundamental frequency contours for each segment estimated by the F0 stylisation algorithm in order to build a discrete model for each speaker.

Within the framework of the SuperSID project, Peskin et al. (2003) conducted an experiment combining both contour statistics modeling and overall distribution statistics. The experiment used statistical values collected on conversation-sides in the Switchboard-I corpus. Several prosodic characteristics were captured by obtaining a vector of N features related mainly to fundamental frequency, segment duration and pauses. For each individual feature, the mean and standard deviation over all words were computed.

In this new approach, each feature vector —i.e. each conversation side's statistics— was viewed as a point in an N-dimensional feature space, and the T training conversations per target speaker formed a *cloud* of T such points. Then, given a test conversation, the distance of its corresponding point to this target cloud was computed, and a k-nearest neighbour classifier was used after comparing the distance to the target speaker cloud versus the average distance to the data clouds for a collection of cohort speakers. The test-trial score was the log likelihood ratio of target and average cohort distances.

Contour matching

In contour matching based systems, some distance is calculated between the contours of prosodic features to recognise speakers. Doddington (1971) and Lummis (1973) are some representative works of this modeling class, where the contour of pitch, intensity and formant frequencies are used to perform speaker recognition.

However, these works are limited by the fact that the spoken message must be the same for training and for testing. In order to overcome this requirement, the work of Adami et al. (2003) proposes a new approach based on the fundamental frequency contour template matching of frequent words (*right*, *okay*, *well*, *uhhuh*, *true*, *really*, *like*, *sure*, *yeah*, *absolutely*, *I mean*, *I know*, *you know*, *I think* and *I see*) extracted from the Switchboard-I corpus (Godfrey et al., 1990). These words seem to have a low dependency on the context and, hence, fairly speaker-specific. Using a dynamic time warping algorithm, the distance between the utterances of the same words or phrases in the test conversation and the templates is computed.

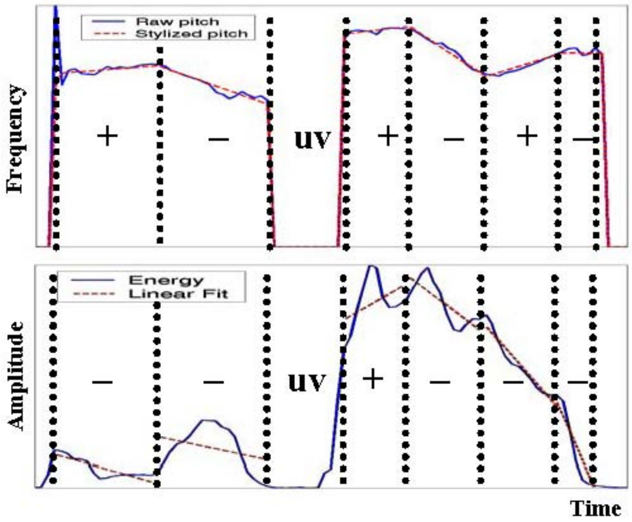


Figure 2.16: State symbol sequence estimation for F0 and energy contours. (After Adami et al. (2003)).

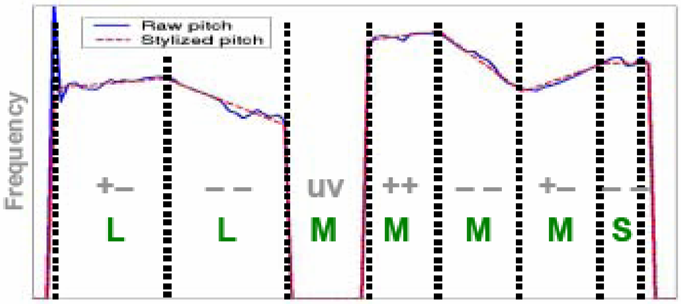


Figure 2.17: Symbol sequence for F0 and energy contours tagged with quantised duration: short (S), medium (M) and long (L). (After Reynolds et al. (2002)).

Based on the approach according to which F0 and intensity have a high degree of interdependence (Lehiste, 1970; Atkinson, 1978; Boves and Strik, 1988; Werner and Keller, 1994), and that they can jointly represent certain prosodic gestures that are characteristic of a particular speaker, Adami et al. (2003) use bigram models to model the dynamics of the fundamental frequency and energy trajectories for each speaker. These trajectories are represented by converting fundamental frequency and energy contours into a sequence of tokens reflecting the joint state of the contours: rising (+) or falling (−); such states are determined using the F0 piecewise linear stylisation algorithm provided in the SRI database (see Figure 2.16).

Information about sound duration is also integrated in the symbol sequence in order to provide a better characterisation of the speaking style of the speaker (Figure 2.17). Furthermore, additional information is added to the contour dynamics by conditioning them on the phoneme context in which they occur (Figure 2.18).

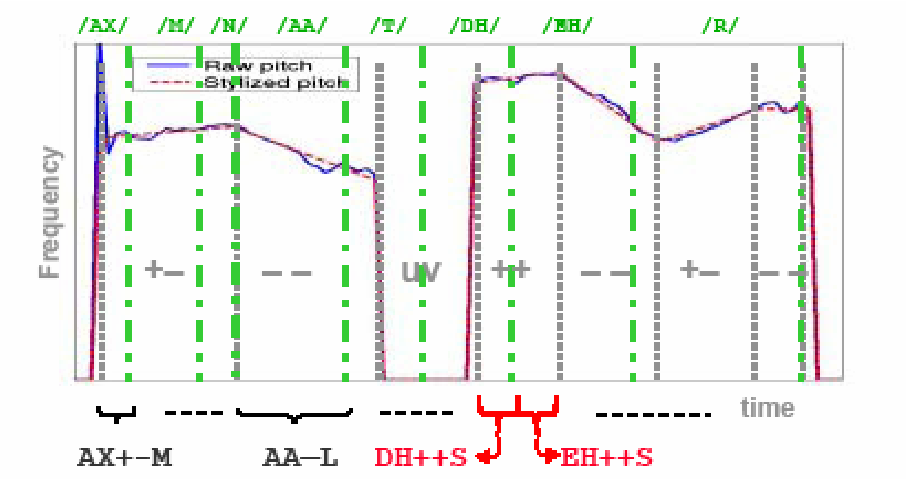


Figure 2.18: Phoneme context information added to F0 slope and duration features. (After Reynolds et al. (2002)).

The work of Adami et al. (2003) has been recently improved in Adami (2007) by using the rate of change of F0 and short-term energy contours. In these recent experiments, delta features are used in place of the F0 stylisation algorithm to model —by means of n-grams— the joint state of the contours. In Adami (2005), the same author describes a system that models the dynamics of the fundamental frequency and temporal trajectories from frequency bands to characterise speaker information. In this system, only the bands that carry most of the relevant information (phone and speaker) are used, and the irrelevant channel information is discarded.

2.4 Multimodal fusion

A biometric modality is the biometric characteristic used in a biometric process. In section 2.1.1, some of the most used biometric modalities —or biometric identifiers— were reviewed, focusing on their main characteristic and possible applications.

This chapter deals with multimodal biometric fusion, which is the combination of information from multiple modalities (voice, face, fingerprints, iris, hand geometry, etc). The aim of multimodal biometric fusion is to obtain better results than using unimodal biometric recognition —the use of a single biometric modality— (Bolle et al., 2004), since some biometric applications require a level of technical performance that is difficult to achieve with a single biometric measure. Moreover, the use of multiple different biometric modalities helps to reduce risk in applications for national identity cards, for example, or security checks for air travel. Multimodal applications are also needed for people who are unable to give a reliable biometric sample for some biometric modalities.

There are several levels at which fusion can take place in a multimodal biometric recognition system. As a basis for the definition of these levels of fusion, a single biometric process is first introduced, using the example of a verification system. Figure 2.19 shows the block diagram of this recognition system based on one single modality.

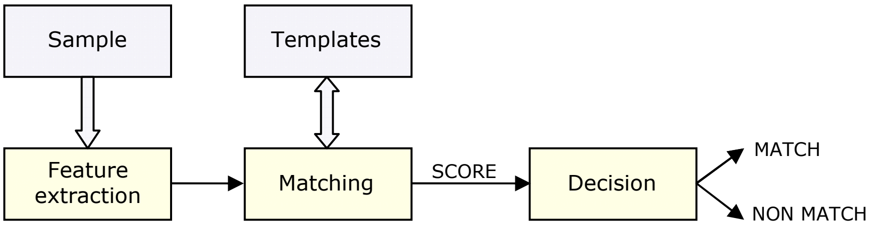


Figure 2.19: Biometric verification process using one single biometric.

In this process, a biometric sample captured by a biometric sensor is fed into the feature extraction module. The feature extraction module converts the sample into features (e.g. fundamental frequency for a voice sample, or minutiae for a fingerprint sample), so that the features are a suitable representation for matching. The features are normally collected into a feature vector; then, the matching module takes this vector as input and compares it to a stored template. The result is a match score, which is used by the decision module to decide —by applying a threshold—, whether the input sample matches with the stored template.

As it was said above, combination of multiple modalities is possible at different levels. These include fusion at the three levels represented by the different blocks in the diagram (Figure 2.19): the *feature extraction level*, the *match score level* and the *decision level*.

Fusion at the feature extraction level occurs before the matching module is invoked. After each individual biometric system has output a collection of features, this combination process fuses all these collections into a single feature set or vector. Score-level fusion combines the individual scores obtained by each individual biometric process into a single score. Finally, fusion at the decision level performs logical operations upon the unimodal system decisions to reach a final resolution. This fusion process fuses the outputs of each individual biometric process, which are represented as Boolean results, using a combination algorithm. Next, more detailed characteristics and the state of the art in each of the mentioned levels of multimodal biometric combination is briefly reviewed.

2.4.1 Feature-level fusion

In feature-level fusion, biometric information is combined after performing the feature extraction. The features extracted from different unimodal systems can be combined in several ways. The simplest form is to combine the feature vectors of the modalities involved and to apply feature classification methods to the combined feature vector.

When a dependence between features from the different modalities exists, fusion at the feature level should allow these dependencies to be more fully exploited than by using only fusion at the score level, so that a better overall performance would be achieved. Nevertheless, fusion at the feature level is difficult to achieve due to the following reasons (Ross and Govindarajan, 2005):

- the feature vectors of multiple modalities may be incompatible (e.g. minutiae set of fingerprints and Eigen-coefficients of face);
- the relationship between the feature spaces of different biometric systems may not be known;
- concatenating two feature vectors may result in a feature vector with very large dimensionality leading to the curse of dimensionality (Bellman, 1961); and
- a significantly more complex matcher might be required in order to operate in the concatenated feature vector.

Fusion at the feature level can be found in literature in several contexts. Chang et al. (2003) use feature-level fusion and ear biometric identifiers achieving significant improvements in performance. On the other hand, Kumar et al. (2003) integrate the palm-print and hand geometry features of an individual enhancing matching performance (in their experiments, fusion at the match score level was observed to be better than fusion at the feature level). However, Ross and Govindarajan (Ross and Govindarajan, 2005) combine the hand and face modalities of a user as well as the R, G, B channels of the face image of a user at the feature level and demonstrate that a feature selection scheme may be necessary to improve matching performance at this level. Therefore, an appropriate feature selection scheme must be used when combining multimodal information at the feature level.

Features can also be combined in a more complex way on an algorithmic level by using other techniques. Since feature extraction algorithms usually require the localisation of landmarks in order to establish a common coordinate frame between samples for feature extraction, components of multimodal biometric systems can exchange these landmarks or mutually their extraction. For example, a face recognition algorithm may provide eye locations for an iris recognition algorithm, or depth landmarks in a 3D face recognition system may be used to correct the pose of faces in texture images. This technique is called co-registration and it is also considered a form of feature-level combination.

2.4.2 Score-level fusion

Each single unimodal recognition process provides matching scores indicating the similarity of the feature vector with the template model or vector. Score-level fusion combines these obtained scores in order to improve the overall performance of the system.

From a theoretical point of view, unimodal biometric systems can be combined reliably guaranteeing improvement in performance at the matching module. The matching scores of any number of such unimodal processes can be combined in a proper way, so that the multimodal biometric combination is, on average, guaranteed to be no worse than the best of the individual biometric systems. The key issue is to identify correctly the fusion method that will combine all these scores in a reliable way, maximising the improvement in matching performance.

Fusion at the match score level is usually preferred by most of the systems. Combining results at the score level typically requires knowledge of both genuine and impostor distributions. However, all of these measures are highly application dependent and generally unknown in any real system; therefore, research on the methods not requiring previous knowledge of the score distribution is still progressing.

Prior to combining the scores of the matchers into a single score, a normalisation process needs to be performed in order to transform all the scores of the individual matchers into a common domain. Therefore, score combination is, in fact, a two-step process: normalisation and fusion itself (Fox et al., 2003; Indovina et al., 2003; Lucey and Chen, 2003; Wang et al., 2004).

Score normalisation

Unimodal scores are usually non-homogeneous; therefore, score normalisation attempts to transform the different scores of each unimodal system into a comparable range of values. As it was reported in Jain et al. (2005), some issues need to be considered prior to fusing the different unimodal scores:

1. First of all, the matching scores at the output of the individual matchers may not be homogeneous. As the example given in Jain et al. (2005), one matcher may output a distance (dissimilarity) measure while another may output a proximity (similarity) measure.

2. Second, the outputs of the individual matchers need not be on the same numerical scale (range).

3. Finally, the matching scores at the output of the individual matchers may follow different statistical distributions.

For all these reasons, it is essential to transform the scores of the individual matchers into a common domain before the score fusion step. The diagram in Figure 2.20 shows a biometric verification process including normalisation and fusion at the score level of two different modalities.

A good normalisation technique must deal with two issues: *robustness* and *efficiency* (Jain et al., 2005). Both properties must be found in the estimates of the location and scale parameters of the matching score distribution, where robust means being insensitive to the presence of outliers and efficiency refers to the proximity of the obtained estimate to the optimal estimate when the distribution of the data is known (Huber, 1981). A lot of methods can be used to normalise a set of scores, but the point is to identify an appropriate robust and efficient technique. Next, the most commonly used score normalisation methods are briefly described and listed in Table 2.4.

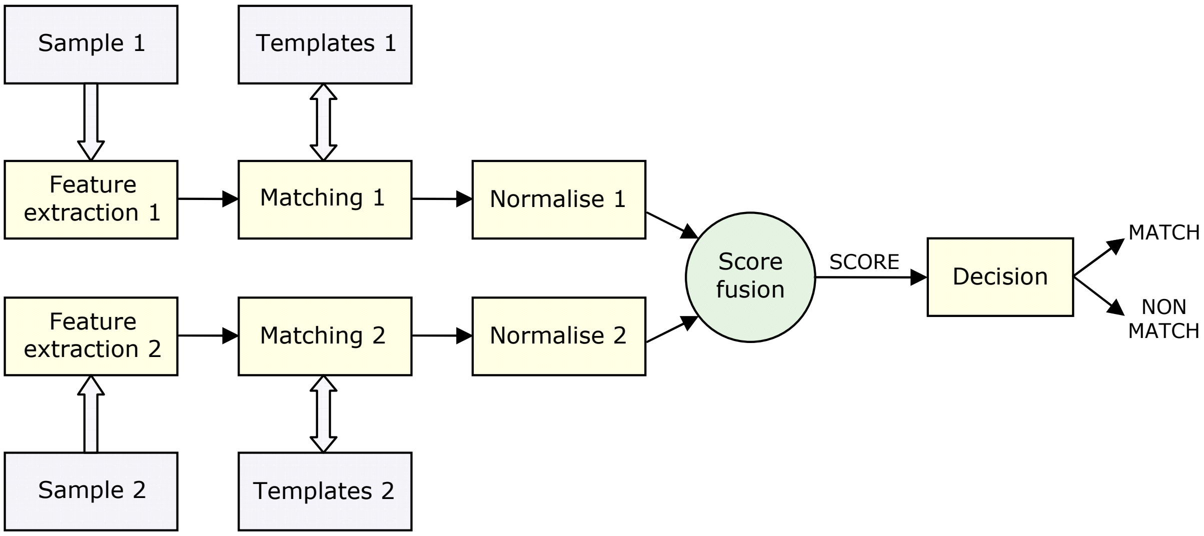


Figure 2.20: Verification process using fusion of two biometric modalities at the score level.

Min-max is the simplest score normalisation technique, in which the minimum and the maximum scores are shifted to 0 and 1, respectively, according to the following expression:

$$s_{MM} = \frac{a - \min(A)}{\max(A) - \min(A)} \tag{2.19}$$

where a is a raw matching score from the set A of the original unimodal biometric scores. This normalisation is a good option when these extreme score values are known, although they can be estimated for a set of matching scores before applying the normalisation if they are not known. When the minimum and maximum values are estimated from the given set of matching scores, the method has the drawback of being not robust, since it is highly sensitive to outliers in the data used for estimation.

Z-score is one of the most conventional score normalisation methods, which normalises the global mean and variance of the scores of a unimodal biometric. Denoting a raw matching score as a from the set A of all the original unimodal biometric scores, the z-score normalised biometric s_{ZS} is computed according to:

$$s_{ZS} = \frac{a - \text{mean}(A)}{\text{std}(A)} \tag{2.20}$$

where $mean(A)$ is the statistical mean of A and $std(A)$ is the standard deviation. The z-score normalisation sets the mean of the normalised scores s_{ZS} to zero, and their variance to one. As reported in Jain et al. (2005), z-score normalisation does not guarantee a common numerical range for the normalised scores of the different matchers. This method is optimal for Gaussian data, where mean and standard deviation are the optimal location and scale parameters; for an arbitrary distribution, both mean and standard deviation are reasonable estimates of location and scale, respectively, but they are not optimal. Hence, if the input scores are not Gaussian distributed, z-score does not retain the input distribution at the output. As the min-max normalisation, the method is also lacking in robustness, since both mean and standard deviation are sensitive to outliers.

Decimal scaling method can be applied when the scores of different matchers are on a logarithmic scale (Jain et al., 2005). Denoting a a raw matching score from the set A of all the original unimodal biometric scores, the normalised score is given by:

$$s_{DSc} = \frac{a}{10^j} \tag{2.21}$$

where j is the smallest integer such that $max|A| \leq 1$. The drawbacks of this method are also the fact of being sensitive to outliers and the assumption that that the scores of different matchers vary by a logarithmic factor.

In contrast to the previous normalisation methods, the **median** and **median absolute deviation** (MAD) method provides a robust normalisation, since both measures are insensitive to outliers and the points in the extreme tails of the distribution (Jain et al., 2005). The normalised score using median and MAD is expressed as:

$$s_{MAD} = \frac{a - median(A)}{MAD(A)} \tag{2.22}$$

where $MAD(A) = median(|a - median(A)|)$. When the score distribution is not Gaussian, median and MAD are poor estimates of the location and scale parameters; hence, the normalisation method has a low efficiency compared to the mean and the standard deviation estimators. As in the z-score normalisation, this thechnique does not retain the input distribution and does not transform the scores into a common numerical range.

Double sigmoid is a normalisation technique similar to min-max normalisation followed by the application of two-quadrics (QQ) or logistic (LG) function (Jain et al., 2005; Snelick et al., 2005). The normalised score is given by:

$$s_{DSi} = \begin{cases} \frac{1}{1+\exp\left(-2(a-t)/r_1\right)} & \text{if } a < t \\ \frac{1}{1+\exp\left(-2(a-t)/r_2\right)} & \text{otherwise;} \end{cases} \tag{2.23}$$

where t is the reference operating point, and r_1 and r_2 denote the left and right edges of the region in which the function is linear. This normalisation method transforms all the scores into a common interval $[0, 1]$, but the shape of the original distribution is not retained. In order to obtain a good efficiency, the parameters t, r_1 and $_2$ must be tuned in a proper way. When the parameters are appropriate, this method is robust and highly efficient. Figure 2.21 shows an example of a double sigmoid normalisation using some fixed parameters ($t = 200$, $r_1 = 20$ and $r_2 = 30$).

In **tanh-estimators**, the normalised score is given by the following expression:

$$s_{TE} = \frac{1}{2}\left\{ \tanh\left(0.01\left(\frac{a - mean_H(A)}{std_H(A)}\right)\right) + 1\right\} \tag{2.24}$$

where $mean_H$ and std_H are the mean and standard deviation estimates of the unimodal score distribution A given by the Hampel estimators (Brunelli and Falavigna, 1995; Snelick et al.,

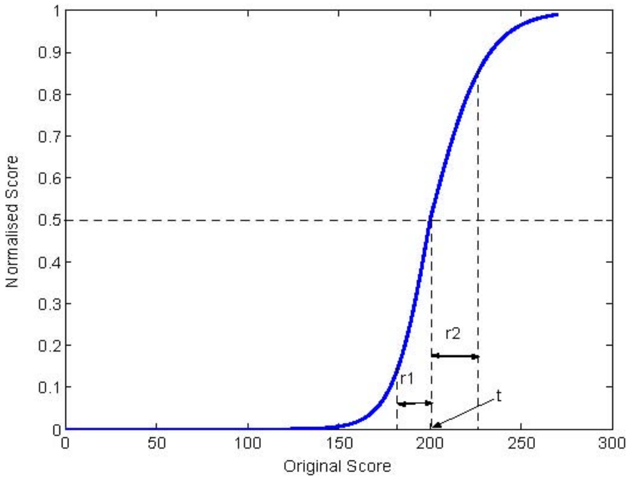


Figure 2.21: Example of a double sigmoid normalisation wit $t = 200$, $r_1 = 20$ and $r_2 = 30$. (After Jain et al. (2005)).

2003). Nevertheless, Jain et al. (Jain et al., 2005) observed that, in their experiments, considering the mean and standard deviation of only genuine scores resulted in a better recognition rate. This method is not sensitive to outliers, since the influence of the points at the tails of the distribution is reduced. Hence, choosing appropriately the parameters used to compute the Hampel estimators, this normalisation method can show a high robustness and efficiency.

Norm. method	Formula	Characteristics
Min-max	$s_{MM} = \dfrac{a - \min(A)}{\max(A) - \min(A)}$	Uses empirical data No accounting for non-linearity Sensitive to outliers
Z-score	$s_{ZS} = \dfrac{a - \mathrm{mean}(A)}{\mathrm{std}(A)}$	Assumes normal distribution Symmetric about mean Sensitive to outliers
Decimal scaling	$s_{DSc} = \dfrac{a}{10^j}; \quad j : max\|A\| \leq 1$	Assumes logarithmic variation Sensitive to outliers
Median and MAD	$s_{MAD} = \frac{a - median(A)}{MAD(A)}$ $MAD(A) = median(\|a - median(A)\|)$	Assumes logarithmic variation Sensitive to outliers
Double sigmoid	$s_{DSi} = \begin{cases} \frac{1}{1+\exp(-2(a-t)/r_1)} & \text{if } a < t \\ \frac{1}{1+\exp(-2(a-t)/r_2)} & \text{otherwise;} \end{cases}$	Highly efficient Insensitive to outliers
Tanh-estimators	$s_{TE} = \frac{1}{2}\left\{ \tanh\left(0.01\left(\frac{a - mean_H(A)}{std_H(A)}\right)\right) + 1\right\}$	Highly efficient Insensitive to outliers

Table 2.4: Some of the most commonly used score normalisation methods.

Histogram equalisation is a general non-parametric method used to match the cumulative distribution function of some given data to a reference distribution. This method was first designed for the enhancement of images, and it was further used for the speech recognition adaptation approaches and the correction of non-linear effects typically introduced by speech systems (Balchandran and Mammone, 1998; Hilger and Ney, 2001). The objective of this method is, in short, to find a non-linear transformation to reduce the mismatch of the statistics of two signals. In Pelecanos and Sridharan (2001) and Skosan and Mashao (2006), this concept was applied to the acoustic features in order to improve the robustness of a speaker verification system. In the current thesis, histogram equalisation will be applied to the scores.

Score fusion techniques

Fusion at the match score level is also known as fusion at the *measurement level* or *confidence level*. In absence of feature-level information, the match score generated by a matcher contains the richest information about the input biometric sample. Moreover, it is relative easy to access and combine the output scores extracted from several matchers. Therefore, fusion at the score level is the most common approach when fusing several biometric modalities.

In the context of a verification task, two distinct approaches to score-level fusion can be considered: the *combination* approach and the *classification* approach (Jain et al., 2005). The first approach formulates the score fusion as a combination problem. Methods used in this approach are some of the simplest and most effective, provided scores are homogeneous or previously normalised. The most popular techniques are sum, product, max, min and median rules, which use simple arithmetic or rule operations to combine scores from multiple sources. In these methods, the individual matching scores are combined to generate a single scalar score, which is then used to make the final decision. Some of the most common fusion methods based on this approach are described further on and summarised in Table 2.5.

Simple sum —also known as *sum rule*—, is the most straightforward fusion method based on the combination approach. All the scores (s_i) of the N normalised unimodal biometric systems are directly summed, resulting in a multimodal final score u_{SS}:

$$u_{SS} = \sum_{i=1}^{N} s_i \tag{2.25}$$

In the **product rule**, the resulting score u_P is obtained by multiplying the normalised scores of the individual biometrics:

$$u_P = \prod_{i=1}^{N} s_i \tag{2.26}$$

Other fusion rules are based on the extreme values of the unimodal biometrics. **Max rule**, for example, takes the maximum normalised score from all the N unimodal biometrics as the final score, while **min rule** takes the minimum value:

$$u_{Max} = \max_{i=1}^{N} (s_i); \qquad u_{Min} = \min_{i=1}^{N} (s_i). \tag{2.27}$$

In **matcher weighting** fusion —also known as *weighted arithmetical mean*— each unimodal score is weighted by a factor proportional to its recognition rate, so that the weights for more accurate matchers are higher than those for less accurate matchers. When using the Equal Error Rates (EER), for example, the weighting factor for every biometric is proportional to the inverse of its EER. Denoting w_i and e_i the weighting factor and the EER for the ith biometric, respectively, s_i the individual score of such biometric, and N the number of biometrics, the final score u_{MW} is expressed as (Indovina et al., 2003):

$$u_{MW} = \sum_{i=1}^{N} w_i s_i; \quad \text{where} \quad w_i = \frac{\frac{1}{e_i}}{\sum_{i=1}^{N} \frac{1}{e_i}}. \tag{2.28}$$

In **user weighting** fusion, these different weighting factors are assigned to each user, since some biometric traits cannot be reliably obtained from a small segment of the population (Jain and Ross, 2002; Indovina et al., 2003; Jain et al., 2005).

Fusion technique	Formula
Simple sum	$u_{SS} = \sum_{i=1}^{N} s_i$
Product rule	$u_P = \prod_{i=1}^{N} s_i$
Max rule	$u_{Max} = \max_{i=1}^{N} (s_i)$
Min rule	$u_{Min} = \min_{i=1}^{N} (s_i)$
Matcher weighting	$u_{MW} = \sum_{i=1}^{N} w_{mi} s_i$ (w_{mi}=ith matcher weight factor)
User weighting	$u_{UW} = \sum_{i=1}^{N} w_{ui} s_i$ (w_{ui}=ith user weight factor)

Table 2.5: Some of the most commonly used score fusion methods based on the combination approach.

In the classification approach, a feature vector is constructed using the matching scores of the individual biometric systems; this feature vector is then classified into one of two classes: *accepted* when the user is a client of the system or *rejected* when the user is classified as an impostor. The classifiers used in this approach are typically support vector machines, neural networks, decision trees, k-nearest neighbours, etc. In contrast to the combination approach, these classifiers are capable of learning the decision boundary irrespective of how the vector is generated, so that the scores of the different modalities can be non-homogeneous and no normalisation is required prior to using the classifier in the fusion process.

Support vector machines (SVM) is a state-of-the-art binary classifier and one of the most currently used fusion techniques based on the classification approach. Recent works on statistical machine learning have shown the advantages of discriminitative classifiers like SVM in a range of applications (Hearst, 1998; Cristianini and Shawe-Taylor, 2000).

The SVM algorithm constructs models that contain a large class of neural nets, radial basis function nets and polynomial classifiers as special cases. The algorithm is simple enough to be analysed mathematically, since it can be shown to correspond to a linear method in a high-dimensional feature space non-linearly related to the input space. Given a linearly separable two-class training data, the SVM algorithm finds an optimal hyperplane that splits input data in two classes, maximising the distance of the hyperplane to the nearest data points of each class.

However, data are normally not linearly separable. In this case, non-linear decision functions are needed, and an extension to non-linear boundaries is achieved by using specific functions called kernel functions (Burges, 1998). The kernel functions map the data of the input space to a higher dimensional space —the *feature space*—, by a non-linear transformation. The optimal hyperplane is then constructed in the feature space, creating a non-linear boundary in the input space. The mentioned hyperplane for a non-linearly separable data is defined by:

$$f(x) = \sum_{i=1}^{N} \alpha_i t_i K(x, x_i) + b \tag{2.29}$$

where t_i are labels, K is a chosen kernel and the coefficients α_i are such that the following condition is satisfied:

$$\sum_{i=1}^{N} \alpha_i t_i = 0; \qquad 0 \leq \alpha_i \leq C. \tag{2.30}$$

The vectors x_i are the *support vectors*, which determine the optimal separating hyperplane and correspond to the points of each class that are the closest to the separating hyperplane. N is the number of support vectors and C is an adjustable parameter that controls the effect of the misclassified data.

2.4.3 Decision-level fusion

After a match decision has been carried out for each individual biometric system, these individual decisions can be fused in order to obtain a final decision for the multimodal system. Fusion at the decision level is based on the binary result values match and non-match output by the decision modules.

For multimodal biometric systems consisting of a small number of different modalities, logical values are normally assigned to match outcomes so that fusion rules can be formulated as logical functions. For biometric systems consisting of many unimodal systems, fusion rules are established by voting schemes, the most common of which is majority voting rule. Table 2.6 shows the behaviour of two specific examples of voting schemes: the AND and OR logical functions, which are the two most widely used functions for biometric systems composed of a small number of modalities. The example in the table is given for a case of two biometric modalities.

Decision biometrics 1	**Decision biometrics 2**	**AND-fused decision**	**OR-fused decision**
match	match	match	match
match	non-match	non-match	match
non-match	match	non-match	match
non-match	non-match	non-match	non-match

Table 2.6: AND an OR fusion of decisions for a case of two biometric modalities.

A more advanced fusion strategy at the decision level is based upon individual accept and reject decisions for each sample. Advanced decision-level fusion methods can be classified into two different groups: the *layered* and the *cascaded*. A layered system uses individual biometric scores in order to determine the pass or fail thresholds for other biometric data processing. On the other hand, cascaded systems use pass and fail thresholds of modality-specific biometric samples to determine if additional biometric samples from other modalities are required to reach an overall system decision.

2.5 Voice imitation and conversion

Unlike some biometric identifiers such as fingerprints, DNA or iris, which are highly permanent, voice is characterised by a high degree of variability due to several non-deliberate factors such as aging, intoxication, illness or emotional stress (see section 2.2.2). Moreover, one's voice can be also deliberately modified, such as speaking in falsetto or feigning a speech defect or foreign accent. All these voice changes are manifested as *voice disguise*, which is defined by Rodman (1998) as

> any alteration, distortion or deviation from the normal voice, irrespective of the cause.

Deliberate modifications vary across the speakers. A study reported in Künzel (2000), for example, showed sex-related differences in the strategies employed for disguise by men and women. Nevertheless, since voice disguise is irrespective of the cause, it can also be classified into electronic and non-electronic, depending on whether electronic devices are used or not to alter the voices. Based on these two dimensions, Rodman (1998) introduced a disguise classification, which is shown in the following table together with some common examples:

	Electronic	**Non-electronic**
Deliberate	Electronic manipulation, voice conversion	Voice imitation, use of dialect and foreign accent, falsetto
Non-deliberate	Channel distortions	Hoarseness, illness, intoxication, emotional stress

Table 2.7: Types of voice disguise based on Rodman's classification (Rodman, 1998).

This section will focus on the deliberate (non-electronic and electronic) cells in Table 2.7. More specifically, it will analyse the main characteristics of those disguises related to a sort of mimicry —both in an electronic and non-electronic manner— and their influence in the speaker recognition task. Section 2.5.1 deals with deliberate non-electronic disguises such as voice imitation, use of dialect and foreign accent, etc.; on the other hand, section 2.5.2 will briefly introduce the use of a specific deliberate-electronic alteration: the electronic conversion of voices in order to resemble a chosen target voice.

2.5.1 Voice imitation

Voice imitation is an innate behaviour that can be find in three main areas as a characteristic of communication (see Table 2.8): language acquisition, impersonation for entertainment, and voice disguise for concealing a personal identity (Zetterholm, 2003).

Imitation in language acquisition is used mainly for learning both mother and foreign languages, but also for accommodation of the speaking manner in a community. Changing the own dialect or sociolect, for example, to the ones spoken by a community, can be seen as a way of similarity and integration in a social group. According to Markham (1997), imitation —in the area of acquisition— can be manifested in several ways: repetition of words, reproduction of syntactic structures, phonetic reproduction, etc., being phonological and phonetic acquisition the most clearly imitative processes.

When imitation has the aim of reproducing another speaker's voice and speech behaviour, it is usually called impersonation (Markham, 1997). Impersonators are normally found in entertainment environments, and they have the ability to pretend successfully to be someone else, being able to identify, select and imitate characteristic features of the target speaker. For entertaining taking place on stage, the impersonator normally copies the body language and other non-vocal features of the target person as a complement to the vocal imitation. On the other hand, when the impersonator is not seen by the audience, it is more important to focus on imitation of vocal features; but wherever the impersonation takes place, the imitator normally tends to focus on the most prominent features and to exaggerate them (Zetterholm, 2003).

Voice imitation can also be used to occult one's identity. Such a disguise is normally performed by physical changes in the vocal tract, in order to modify the pitch, voice quality, dialect, accent, prosodic pattern, etc. In this case, the person may simply try to disguise the own voice, and not to imitate any specific person; however, the features changed (pitch, prosody, etc.) are not exaggerated and they could correspond to real features. In the same way, the adopted accent and dialect could exist as well in some region or social group. Therefore, the alteration of both dialect and accent aiming to be alike some existing one is usually referred to as dialect and accent imitation.

Area	**Aim**	**Characteristics**
Language acquisition	First and second languages Speaking manner adaptation	Imitation of several real speakers Repetition of words Reproduction of syntactic structures Phonological/phonetic acquisition
Impersonation	Entertainment	Imitation of a specific person Vocal and non-vocal imitation Exaggerations
Voice disguise	Hide identity	Imitation of a fictitious person No exaggerations

Table 2.8: Voice imitation areas, aims and characteristics.

The analysis of voice disguise is of interest in research to define acoustic features for speaker recognition. In this sense, voice imitation can be used as a method to find out which speech features a person can change with success, and which features are more difficult to alter or imitate. This complementary knowledge will certainly improve the speaker recognition task.

Next, the influence of voice imitation on the human voice and the easiness of speech features of being imitated or modified is outlined. Then, the state of the art in dialect and accent imitation and the robustness of speaker recognition systems to voice imitation are briefly reviewed.

Imitated features

An imitator must focus on important features and passages in the text. When the aim of the imitation is to entertain, the impersonator may exaggerate the most characteristic features in order to convince the audience. Moreover, the impersonator may also use characteristic words or phrases for a more convincing effect.

The variety of features to imitate is great: some features are related to the regional and social environment —such as dialect and sociolect— whereas other features are more individual —such as voice quality, speech style and phonetic habits. Therefore, a successful imitator must consider both features of group identity and individual features, such as person's phonetic habits (Zetterholm, 2003).

However, there are some organic differences between speakers that cannot be changed. When these differences are large, it may be difficult to achieve good imitations of another person's voice (Laver, 1994). An extreme case is found between male and female voices: both voices show differences concerning fundamental frequency, intensity, the shape of glottal wave, etc. (Pittam, 1994). In a study by Lass et al. (1982), some speakers were asked to attempt to speak like the opposite sex, but an auditory analysis revealed the actual sex of the speakers. A professional Swedish imitator interviewed in Zetterholm (2003) stated also the impossibility to imitate female voices. The same imitator declared that he found much easier to imitate older voices than younger voices.

The question arises as whether it is enough to pick out and copy a number of specific voice features from another person, or having a similar voice is more important than the feature selection itself. Some of the most extensive studies about imitations are found in the works of Zetterholm (Zetterholm, 2002a, 2003, 2006), where the influence of prosody (Zetterholm, 2000, 2002b) and semantic information (Zetterholm et al., 2002) is emphasised in some of them. Zetterholm's research raises some questions related to the ability and the way how impersonators select and imitate other acoustic features, for example:

1. How and to what extent does an impersonator change his own voice and speech behaviour?
2. What features are important for successful voice imitation? Are some features more important than others?
3. Which features are hard to imitate?
4. Do different impersonators select and try to imitate the same features when imitating the same target speaker?

The research study involves three male professional impersonators specialised in imitating well-known politicians and TV personalities, and it is mainly performed in terms of fundamental frequency, voice quality, prosodic patterns and dialect, with posterior auditory and acoustic analyses. The experiments reveal that all these features seem to be important for a convincing imitation. Moreover, the imitators tend to change the formant frequencies in their imitations, especially those of the stressed vowels.

In order to know the importance of being a professional imitator, these experiments also included a comparison of two professionals and one amateur imitating the same target voices.

The results showed that, although the voice imitations made by the three impersonators were different in terms of acoustic characteristics and perception, in all cases it was possible to make a clear identification of the imitated person, and the impersonators selected the same and most prominent features of the target voices, focusing mainly on pitch, dialect, rhythm, pausing and phonetic pronunciation of specific segments as well as individual habits —such as hesitations and loud breathing— of the target speakers.

Moreover, in accented words and phrases, the impersonators succeeded in imitating vowels and other specific features, whereas imitators' own voices and dialects were sometimes audible in unstressed passages. Voice quality was also captured in the imitations, although imitating different dialects and individual features —such as hesitations— seemed to be easier than copying different individual voice qualities. The mean F0 was also captured to some extent, since the mean F0 of the impersonators' own voices was always reflected in the imitations. Only a significant difference was found between the professional and amateur impersonators: whereas the professional ones copied successfully the intonation pattern of the target speakers, the amateur one didn't succeed in doing it, failing in maintaining the correct F0 range over time.

The work of Zetterholm (2003) concludes that impersonators usually capture several aspects of the target voice, and that an imitation may be successful on the whole even if they fail in imitating some features, provided they are successful in imitating the most prominent features. Nevertheless, since the most characteristic features differ between speakers, it seems hard to stipulate a general ranking order of important features.

Dialect imitation in speaker recognition

As it has been pointed out above, dialect and accent —together with their specific phonetic features— are part of the set of voice characteristics that can provide relevant information about a person's identity. One common voice disguise is the modification of the own dialect or accent: both dialect and accent are subject to imitation or alteration in order to avoid or complicate the speaker's identification.

The importance of dialect and accent disguise to speaker identification has been pointed out by researchers such as Shuy (Shuy, 1995). Such changes to the voice can thus affect the outcome of voice identification. A detailed understanding of dialect and accent traits, and how they can be disguised is, therefore, essential to the speaker recognition task under the influence of this kind of imitations.

Some studies are focused on whether theatrical accent modification affects human perception in the process of voice discrimination. Stage dialects are a widely investigated phenomenon (Machlin, 1975; Halloran, 2003) and are used in the theatre and film. These dialect imitations are not made to fool people, but rather to draw people into an imaginary world. A dialect imitation could however be used in a criminal setting and, for this reason, it is relevant to investigate how convincing dialects imitated for theatrical use are.

On the other hand, some studies concern themselves with the speaker and dialect imitation research considering both automatic speaker recognition and human speech perception. Both approaches were used, for instance, in research conducted by Sullivan and Pelecanos (2001) and Zetterholm et al. (2004). The work by Sullivan and Pelecanos showed that the recognition system was capable of classifying the mimic attacks more appropriately than human listeners. The work by Zetterholm et al. found a minimal correlation between the speaker verification system they used and their human listeners in how they judged the 62 imitations in closeness to the target speaker.

2.5.2 Voice conversion

Automatic voice conversion is the modification of a speaker voice —called *source speaker*— in order to make it being perceived as if another speaker —*target speaker*— had uttered it. Given thus two speakers, the aim of a voice conversion system is to determine a transformation function that *converts* the speech of the source speaker (from which usually a complete database is available) into the speech of the target speaker (from which normally few data are available), replacing the physical characteristics of the voice without altering the message contained in the speech (Duxans, 2006; Erro and Moreno, 2006).

Voice conversion applications

Voice conversion technology can be found in several applications, specially in those related to the speech synthesis field. In a voice synthesiser, speech is normally generated by concatenation of selected units from a database, which has been previously recorded by a qualified speaker. The inclusion of a voice conversion system at the output of a speech synthesiser allows the customisation of the system, so that no previous recording of speech from the potential users is needed (Erro and Moreno, 2006). Moreover, conversions can be intra-gender and cross-gender, so that female voices can be converted to male voices and viceversa.

Other voice conversion applications include the use of automatic speech-to-speech translation and foreign language learning. In some cases —for instance, in a conference call with more than two participants— the translation task may require speaker identification by listeners, in order to be able to differentiate speakers by their voices (Duxans, 2006). In foreign languages learning, one of the most difficult tasks is the proper intonation of sentences and pronunciation of non-existing phonemes in the native language. Voice conversion may help in this sense, if the students can listen to their own voices with a proper pronunciation (Mashimo et al., 2001, 2002; Duxans, 2006).

For people with some speech impairment, voice conversion systems may be used to improve intelligibility of their abnormal speech (Hosom et al., 2003). On the other hand, voice conversion can also be used for designing hearing aids appropriated for specific hearing problems; for instance, by transforming speech signals into other frequency ranges for those people who are prevented from hearing specific frequency range sounds (Duxans, 2006).

Voice conversion can also be found in the entertainment field; in a karaoke, for example, helping singers to success in any kind of songs, or in the film dubbing industry, by generating voices of famous actors in any language or voices of actors who are not alive. Some voice conversion applications are summarised in Table 2.9.

Conversion techniques

As it was seen it section 2.2.2, speech signal processing extracts features from the speech signal related to the source and filter processes. Earlier voice recognition systems tended to use solely

Applications	Examples and characteristics
TTS customisation	Only a recorded database from the source speaker is needed Any target voice can be created Intra-gender and cross-gender conversions
Automatic TTS translation	Speaker identification by listeners
Foreign language learning	Help students to get a proper pronuntiation
Medical aids	Improve intelligibility of an abnormal speech Design hearing aids appropriated for specific hearing problems
Entertainment	Voices of famous actors or actors who are not alive Help singers in a karaoke

Table 2.9: Voice conversion applications.

these filter parameters, which related to the physiology of the vocal tract and to the learnt articulatory configurations that shape the specific speech sounds. Since the spectrum of the acoustic wave is altered when the sound wave moves through the vocal tract, these filter parameters are referred to as the spectral level of speech.

On the other hand, human speaker recognition relies on other information sources like a specific use of lexical terms, prosody or phonetics. These linguistic characteristics are highly dependent on the learned habits or style, and they are manifested in the dialect, sociolect or idiolect of the speaker. Nevertheless, the state-of-the art voice conversion systems use only the former —spectral— features in the voice transformation between two speakers. As far as it is known, no prosodic or other linguistic level features have been applied in the voice conversion field (Duxans, 2006).

As in speech coding, speech recognition and speech synthesis, the feature extraction process for voice conversion consists in segmenting the speech signal into (usually) overlapping frames. Each frame is then represented by a series of coefficients, and the dimensionality of the speech samples is reduced. Three different approaches for feature extraction in voice conversion are commonly found in literature (Duxans, 2006), depending on the type of features extracted:

1. **Features related to a phonetic/acoustic model of the speech.** Systems based on features related to a phonetic/acoustic model —such as formant frequencies, formant bandwidths and glottal flow parameters— try to change the voice characteristics by modifying several aspects of the signal. Since phonetic/acoustic parameters are related to the physical speech production, it can be assumed that these systems are able to find specific conversion functions and transform the signal in detail; however, a reliable estimation of these parameters is difficult to achieve.

2. **Features related to Linear Prediction.** Linear Prediction (LP) features are based on the source-filter model for speech production. These features are frequently used in voice conversion systems. Estimation techniques of LP coefficients are more robust than other parametric feature estimations. Morever, LP separates the vocal tract contribution from the excitation contribution, so that each signal component can be converted in a different way. Nevertheless, in LP systems, two different mappings need to be trained: one for the vocal tract parameters and another for the LP residual signal.

3. **Features with no speech model assumption.** Systems using LP related features or features without assuming any signal model —such as spectral lines or mel frequency cepstrum coefficients— utilise techniques for global optimisation and control, instead of modifying each phonetic/acoustic parameter separately.

The type of acoustic features selected influences clearly the final performance of voice conversion systems. Therefore, although it is important to select features that capture the speaker identity, these need to be low dimensional features with good interpolation properties as well. Moreover, since the converted features need to be transformed into speech, a speech production technique based on the selected features must be available.

Some recent concatenative voice synthesis systems are based on the TD-PSOLA algorithm —which stands for *time domain pitch synchronous overlap-add*— in which no speech signal model is assumed, but the synthesis is directly performed from the speech units. These systems are thus not appropriate for voice conversion. Moreover, they are not robust to noise derived from unit concatenation (Erro and Moreno, 2006).

Other systems are based on the LP-PSOLA (*linear prediction pitch synchronous overlap-add*) algorithm have been shown a good performance (Duxans et al., 2006; Sündermann et al., 2006). However, when resynthesising the modified signal, some noise coming from the phase incoherence is introduced. Systems based on harmonic or sinusoidal decomposition of the signal have been successfully used in voice conversion because of their high flexibility (Stylianou et al., 1998; Ye and Young, 2006); however, they are usually pitch-synchronous systems, whose quality is conditioned to the division of the input signal in a very precise way into its fundamental periods (Erro and Moreno, 2006).

Most voice conversion works make use of a parallel training corpus, which contains acoustic feature vectors from both source and target speakers corresponding to the same speech sound (Kain and Macon, 1998; Stylianou et al., 1998). In order to develop such corpus, the same sentences are usually recorded by both source and target speakers, aligning properly the equivalent speech fragments. Nevertheless, recordings of identical sentences from both speakers are not always available; in order to solve this problem, several techniques have been recently proposed (Duxans et al., 2006; Sündermann et al., 2006; Ye and Young, 2006).

2.5.3 System robustness to voice imitation and conversion

Voice imitation and disguise are potential threats to security systems using automatic speaker recognition. Therefore, several studies have been done to test the vulnerability of speaker recognition systems against voice disguise and imitations by human or synthetic voices. An experiment reported in (Lindberg and Blomberg, 1999) tried to deceive a state-of-the-art speaker verification system by using different types of artificial voices created with client speech. Other works related to the vulnerability of automatic recognition systems to specifically created synthetic voices can be found in Masuko et al. (2000) and Matrouf et al. (2006), where the impostor acceptance rate is increased by modifying the voice of an impostor in order to target a specific speaker.

Other studies have dealt with the effects of common types of voice disguise against automatic speaker recognition systems. Some preliminary experiments reported in Künzel et al. (2004) about the effects of increased voice pitch, lowered pitch and pinching the noise while speaking, showed that the performance of an automatic speaker recognition system was degraded by all three modes of disguise, where the mode *lowered pitch* presented the smallest degradation.

The vulnerability of state-of-the-art speaker recognition to human imitations has been tested in some recent studies (Lau et al., 2004, 2005), where the experiments showed that an impostor who knows a client speaker of the database with a similar voice to his own voice, could attack the system. Such vulnerability is of particular concern where automatic speaker recognition is used to control client access in applications such as telephone banking or other financial services. On the other hand, some experiments reported in Sullivan and Pelecanos (2001) showed that an automatic speaker verification system was more robust against an impostor who could closely imitate a target voice than those systems relying on human identification and verification. Other cases include the use of twins voices, which could, in some cases, be considered as a sort of voice imitation. In order to know how they differed, an experiment with twins performed in Scheffer et al. (2004) showed that an automatic speaker recognition system was able to identify a twin with an acceptable performance (85% of correct identification) and that the system was able to discriminate the target speaker from its twin in the verification mode.

Generally, systems that use both source and filter parameters perform better than systems that just use source parameters, when systems are evaluated by means of generic background models and without impostors who employ intentional voice mimicking techniques. Where a speaker recognition system utilises both source and filter parameters, and since prosodic features have been considered for state-of-the-art recognition systems in recent years, the question arises whether either the source or the filter parameters are more vulnerable to intentional mimicking. In Lau et al. (2004), it transpired that the mimicking subjects, both with and without training in phonetics, found it easier to mimic the source parameters of the target speaker than the filter parameters. Another study showed, however, that a professional voice imitator from the entertainment industry was clearly able to approximate the filter parameters of a well-known target speaker (Zetterholm, 2006).

Chapter 3

Using Prosody, Jitter and Shimmer for Speaker Recognition

Prosody, lexis and phonetics are some of the linguistic levels of information used by humans to recognise others by their voice, as it was seen in section 2.2.2. These levels of information are normally related to learned habits and style, and they are mainly manifested in the dialect, sociolect or idiolect of the speaker.

Since these linguistic levels play an important role in the human recognition process, some effort has been placed in adding this kind of information to automatic speaker recognition systems. Some works (Peskin et al., 2003; Reynolds et al., 2003) have recently demonstrated that prosodic features help to improve recognition systems based solely on filter parameters, supplying complementary information not captured in the traditional systems. Moreover, some of these parameters have the advantage of being more robust than spectral features to some common problems like noise, transmission channel distortion, speech level and distance between the speaker and the microphone (Atal, 1972; Carey et al., 1996).

The objective of this chapter is twofold. First, the chapter focuses on the use of prosody in speaker recognition, and the combination of both prosodic and spectral information in order to improve the overall performance of a verification system. The baseline spectral system used in this experimental part is introduced in section 3.1. Section 3.2 includes the description of the prosodic parameters used to improve the baseline system (3.2.1) and their individual performance as speaker discriminant parameters is tested over the conversational Switchboard-I database (3.2.2). Furthermore, several fusion and normalisation techniques are applied to combine prosodic features themselves (section 3.2.3) and with the spectral parameters (section 3.3).

Apart from prosodic features, there are probably many more characteristics that may provide complementary information and should be of a great value for the speaker discrimination task. Jitter and shimmer, for example, are acoustic characteristics of voice signals and they are quantified as the cycle-to-cycle variations of fundamental frequency and waveform amplitude, respectively. Since they have been widely used as detectors of voice pathologies, they seem *a priori* to be useful for discriminating speakers.

The second objective of this chapter is to improve a prosodic and voice spectral verification system by introducing new features based on jitter and shimmer measurements. In section 3.4, both jitter and shimmer are introduced in more detail and several methods to measure such parameters are described. Both features are also used to perform some speaker verification experiments again over the Switchboard-I database. In section 3.5, some selected jitter and shimmer measurements are used in combination with prosodic and short-term spectral parameters. Finally, conclusions of the experiments are presented in section 3.6.

3.1 Spectral baseline system

As it was seen in section 2.2.2, the spectral shape of the speech signal contains information about the vocal tract and the excitation source in the glottis by means of the formants and the fundamental frequency, respectively. Therefore, most of the parameters used in speaker recognition are obtained from the spectrum of the signal. In this section, a verification system based on spectral coefficients is introduced. When using prosodic information, this system will serve as a baseline in order to see how useful and complementary prosodic parameters are.

In the previous section, cepstral coeficients were showed to be the usual way of representing the short-time spectral envelope of a speech frame in current speaker recognition systems. However, such coefficients have some disadvantages that are overcome by using Frequency Filtering parameters (see section 2.2.3). These parameters have been used in our experiments since they give comparable or even better results than mel-cepstrum coefficients in the experiments performed in Nadeu et al. (1995) and Abad et al. (2003).

Thus, the spectrum-based recognition system used in this section is a 32-component GMM-UBM system, using short-term feature vectors consisting of 20 Frequency Filtering parameters with a frame size of 30 ms and a shift of 10 ms. In order to take into account the dynamic information along the frames, 20 corresponding delta and acceleration coefficients have also been included.

All the verification experiments described in this chapter have been performed with the Switchboard-I database (Godfrey et al., 1990), which is a collection of 2430 two-sided telephone conversations among 543 speakers from all areas of the United States. In order to use the full collection of speakers as target speaker, NIST defined a jack-knifed test design, splitting the corpus into six partitions (or *splits*). Speakers within each split can be used as target or impostor speakers for that split, and speakers in the other five splits may be used for training and normalization without fear of speaker contamination. Cycling around all six splits uses effectively the complete corpus (Reynolds et al., 2002).

The NIST Extended Data Task specifies different training conditions, namely 1, 2, 4, 8, or 16 conversation sides for training the target speaker models. In these experiments, the models have been trained with 8 conversation sides, since relatively few speakers in Switchboard-I participated in 16 calls. Splits 1-3 of the Switchboard-I database have been used as a training set for the speaker models, and the UBM has been trained with 116 conversation sides, randomly taken from the corresponding complementary set of cohort speakers (splits 4-6). The system was tested using one conversation-side for each test trial, according to the NIST's 2001 Extended Data Task (Doddington et al., 2000).

By using this experimental setup, the Equal Error Rate obtained in the spectrum-based speaker recognition system equals **10.1**±1.4**%**. The margin error (±1.4) has been calculated by a 95% Wald confidence interval for a proportion according to the following expression (Newcombe, 1998):

$$\varepsilon \simeq 1.96\sqrt{\frac{p(1-p)}{N}} \tag{3.1}$$

where p is the computed proportion (result) and N the number of experiments performed. From this point on —unless otherwise stated— all the margin errors will be computed in this way.

3.2 Prosodic system

At the 2002 Summer Workshop on Language Engineering at the Johns Hopkins University —Center for Language and Speech Processing—, the SuperSID project was undertaken with the aim of improving the accuracy of automatic speaker recognition technology by exploiting new information sources in conversational speech (Peskin et al., 2003; Reynolds et al., 2002). The extraction and modeling of speaker-specific patterns in acoustic cues, speech prosody, word and phone pronunciations, word usage and interactions with conversational partners were explored and developed.

Within the framework of the SuperSID project, Peskin et al. (2003) focused on investigations of a diverse collection of prosodic features —apart from exploring modeling conversational patterns— from the Switchboard-I corpus. The 2001 Extended Data Task was used as a testbed in part because of the well-defined test protocol and well-established baselines for this task, but mainly because of the wide variety of data resources available for the Switchboard-I corpus (Reynolds et al., 2002).

The work on prosodic features was made possible through the availability of the Prosodic Feature Database developed by SRI (Shriberg et al., 2000). This was a new version of an earlier database developed by SRI for applications in topic and sentence segmentation. The original database contained the transcriptions only for a subset of Switchboard-I, and it was also used later by Weber et al. (2002) for research experiments on speaker recognition on a small subset of the Extended Data Task. However, this database was not available for the realisation of this thesis. Instead, prosodic features were extracted using the manually corrected word-level transcriptions of the entire Switchboard-I corpus provided by ISIP (Mississippi State University), and publicly available, along with associated documentation, at http://www.ece.msstate.edu/research/isip/projects/switchboard/.

3.2.1 Description of the prosodic features

The prosodic features extracted for the experimental part of this thesis, inspired by the previous works of Shriberg et al. (2000) and Peskin et al. (2003), are listed below. These include features related to word and segmental duration, fundamental frequency and pauses, and they were extracted using the ISIP's transcriptions and the Praat software for acoustic analysis (Boersma and Weenink, 1992; Boersma, 1993). Logarithmic transformations are performed on some features so that the distribution of values will look more Gaussian.

Features related to word and segment duration

- Logarithm of number of frames per word (based on a 10 ms frame duration), averaged over all words.
- Fraction of voiced frames within each word, averaged over all words.

The number of frames per word was directly extracted from the ISIP's transcriptions. The fraction of voiced frames within each word was computed with the Praat software. Performing a pitch analysis based on a cross-correlation method, this feature is the fraction of locally pitch frames that are analysed as voiced. The usual pitch analysis contains a *path finder* that searches for a smooth path through the local pitch candidates. This path finder is temporarily switched off to determine the fraction of locally voiced frames. A frame is regarded as *locally voiced* if it has a voicing strength above the *voicing threshold* or a local peak below the *silence threshold*. The voicing threshold is the strength of the unvoiced candidate, relative to the maximum possible autocorrelation, and it was set to its standard Praat value 0.45. The silence threshold, which is set to the standard value 0.03, indicates that the frames that do not contain amplitudes above this threshold (relative to the global maximum amplitude), are probably silent (Boersma and Weenink, 1992).

Features related to fundamental frequency

Fundamental frequency is estimated by Praat, performing an acoustic periodicity detection based on a cross-correlation method (optimised for voice analysis) using a Hanning window with a physical length of 40/3 ms and a shift of 10/3 ms. By using these settings, the following F0-related features are extracted:

- Logarithm of mean F0 for each word, averaged over all words.
- Logarithm of maximum F0 for each word, averaged over all words.
- Logarithm of minimum F0 for each word, averaged over all words.
- Logarithm of the range of F0 ($\max F0 - \min F0$) for each word, averaged over all words.
- F0 slope for each word, averaged over all words, computed as:

  $$\frac{\text{last}F0 - \text{first}F0}{\text{number of frames}} \tag{3.2}$$

  being *number of frames* the frames comprised between the first and the last F0 of each word.
- Mean slope of the stylised F0 contour obtained for each word, averaged over all words.

The F0 contour stylisation is a Praat function whose aim is to obtain a much simplified F0 curve, according to the following algorithm (Boersma and Weenink, 1992):

1. Look up the F0 point that is closest to the straight line that connects its two neigbouring points.

2. If this F0 point is further away from that straight line than a *frequency resolution*, the stylisation is finished: the curve cannot be stylised any further.

3. If the stylisation is not finished, the F0 found in step 1 is removed.

4. Go back to step 1.

A frequency resolution of 2 semitones (the standard value given by Praat program) has been used to determine the stylised F0 contour. Figure 3.1 shows an example of a F0 contour stylisation using a 2 semitones frequency resolution.

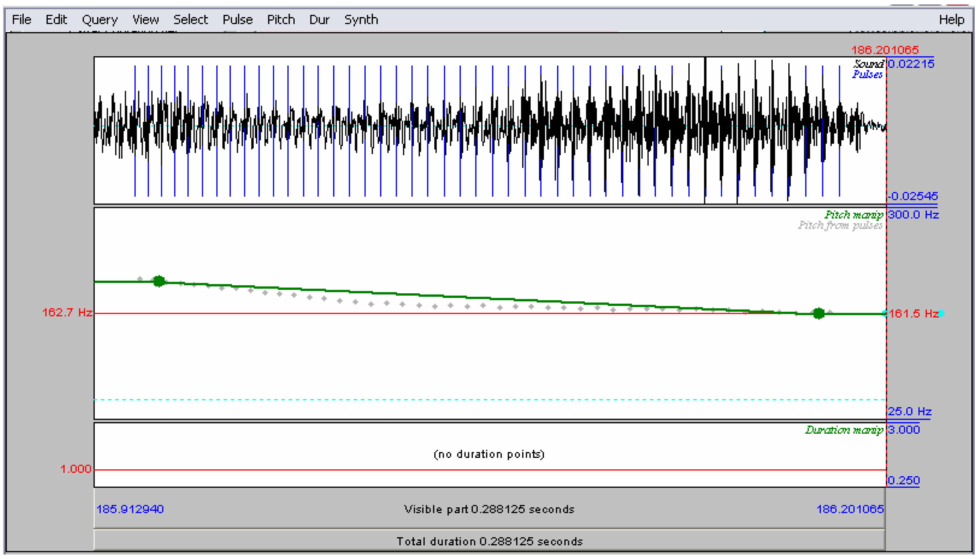


Figure 3.1: Example of a F0 contour stylisation using a frequency resolution of 2 semitones. Green big points indicate the stylised F0 contour and grey small points correspond to the original F0 points.

Features related to pauses

Based on a 10 ms frame duration, pauses were defined according to a classification made by Shriberg et al. (2000), and further used —with different nomenclatures— in Peskin et al. (2003). These include *short pauses* for pauses between 7 and 15 frames length, *medium pauses* between 16 and 99 frames, and *long pauses* for pauses equal to or greater than 100 frames. The pause features extracted for these experiments are:

- Relative frequency of short pauses:

  (silences 7-15 frames long)/(total number of pauses)

- Relative frequency of medium pauses:

  (silences 16-99 frames long)/(total number of pauses)

- Relative frequency of long pauses:

  (silences $\geq$ 100 frames long)/(total number of pauses)

- Logarithm of the short pause duration averaged over all short pauses.
- Logarithm of the medium pauses duration averaged over all medium pauses.
- Logarithm of the long pause duration averaged over all long pauses.

3.2.2 Experiments on individual features

All the verification experiments described in this section were performed with the Switchboard-I database (Godfrey et al., 1990), which is a collection of 2430 two-sided telephone conversations among 543 speakers from all areas of the United States. In order to use the full collection of speakers as target speaker, NIST defined a jack-knifed test design, splitting the corpus into six partitions (or *splits*). Speakers within each split can be used as target or impostor speakers for that split, and speakers in the other five splits may be used for training and normalization without fear of speaker contamination. Cycling around all six splits uses effectively the complete corpus (Reynolds et al., 2002).

A feature vector was obtained by using all 14 characteristics described in the previous section. The system was tested using the k-nearest neighbour classifier in the sum rule approach, comparing the distance of the test feature vector to the k closest vectors of the claimed speaker versus the distance of the test vector to the k closest vectors of the cohort speakers. For those features described as *averages*, the mean and standard deviation over all words were computed for each individual feature; for rates, a flat rate giving relative frequency was simply computed. Then, three distance measures were used: the Euclidean distance, the Mahalanobis distance and the symmetrised Kullback-Leibler divergence expressed as:

$$d_{KL} = \frac{1}{2}(\mu_1 - \mu_2)^2\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right) + \left(\frac{\sigma_1}{\sigma_2} - \frac{\sigma_2}{\sigma_1}\right)^2 \qquad \text{(for stats)} \qquad (3.3)$$

$$d_{KL} = (\rho_1 - \rho_2)\log\left(\frac{\rho_1}{\rho_2}\right) \qquad \text{(for rates)} \qquad (3.4)$$

where μ and ρ are the means and σ the standard deviation.

The NIST Extended Data Task specified different training conditions, namely 1, 2, 4, 8, or 16 conversation sides for training the target speaker models. The models in these experiments were trained with 8 conversation sides, since relatively few speakers in Swithcboard-I participated in 16 calls. In a first step, training was performed using splits 1-3 of Switchboard-I database, and the three held out splits provided the cohort speakers. The system was tested with one-conversation-side according to the NIST's 2001 Extended Data task (Doddington et al., 2000), obtaining 672 client trials and 1188 impostor trials. A second experiment was conducted using splits 4-6 for training and splits 1-3 as cohort speakers. Finally, both previous experiments were combined, using all 6 splits: splits 1-3 with splits 4-6 as cohort set and vice versa, giving a total number of 3724 trials (1343 clients and 2381 impostors).

Tables 3.1 and 3.2 show the EER obtained for each prosodic feature using k=1 and k=3 in the k-nearest neighbour, respectively. Both tables also show the results obtained for each isolated feature using three different distance measurements: Euclidean, Mahalanobis and Kullback-Leibler, and different split sets for training.

Prosodic feature	**Euclidean**			**Mahalanobis**			**Kullback-Leibler**		
	1-3	4-6	1-6	1-3	4-6	1-6	1-3	4-6	1-6
Log (frames/word)	35.7	36.0	35.9	36.0	35.9	36.0	32.9	29.9	31.3
Voiced segments	36.0	37.4	36.6	36.6	37.2	36.9	30.1	32.8	31.3
Log (mean F0)	**26.5**	27.9	27.2	26.8	**28.8**	28.0	20.4	**20.0**	**20.2**
Log (max F0)	26.6	**27.4**	**27.0**	**26.3**	**28.8**	**27.4**	**18.8**	21.6	**20.2**
Log (min F0)	30.6	31.0	30.8	30.6	32.3	31.5	21.9	21.9	21.9
Log (range F0)	34.5	34.7	34.8	35.1	34.9	34.6	27.2	29.4	28.4
F0 slope	40.8	42.6	41.7	39.3	42.3	40.9	39.3	40.0	39.5
Stylised F0 slope	39.8	34.6	37.5	39.0	33.5	36.6	31.0	27.1	29.3
Short pauses frequency	47.6	50.2	48.7	47.6	50.2	48.7	46.8	46.0	46.5
Medium pauses frequency	45.5	43.2	44.4	45.5	43.2	44.4	44.3	43.1	43.5
Long pauses frequency	43.0	44.7	44.1	43.0	44.7	44.1	42.7	44.0	43.3
Log (short pause duration)	49.3	50.2	49.7	49.5	51.4	50.3	48.4	51.3	49.7
Log (medium pause duration)	43.0	43.8	43.4	42.1	44.3	43.2	41.6	42.0	41.8
Log (long pause duration)	47.3	45.7	46.4	48.8	46.9	47.9	42.7	39.5	41.2

Table 3.1: EER (%) for each prosodic feature using different distance measurements and k=1.

Prosodic feature	**Euclidean**			**Mahalanobis**			**Kullback-Leibler**		
	1-3	4-6	1-6	1-3	4-6	1-6	1-3	4-6	1-6
Log (frames/word)	33.8	35.0	34.2	33.4	34.3	34.0	31.6	29.0	30.4
Voiced segments	33.5	34.9	34.6	34.2	36.5	35.3	30.1	32.7	31.6
Log (mean F0)	**24.8**	26.4	25.7	**25.3**	28.8	26.7	**20.4**	**18.2**	**19.2**
Log (max F0)	25.0	**24.6**	**24.8**	25.6	**27.4**	**26.5**	21.0	21.8	21.4
Log (min F0)	27.7	28.3	27.6	28.3	30.3	29.0	22.3	20.7	21.5
Log (range F0)	32.4	33.7	33.1	31.1	34.3	32.2	26.6	28.0	27.2
F0 slope	38.1	41.1	39.7	35.6	40.8	38.5	38.4	39.4	39.1
Stylised F0 slope	37.7	32.8	35.7	38.3	32.5	35.8	29.9	27.6	28.7
Short pauses frequency	46.1	46.8	46.5	46.1	46.8	46.5	46.7	48.7	47.4
Medium pauses frequency	43.7	42.6	43.1	43.7	42.6	43.1	45.0	44.1	44.4
Long pauses frequency	41.2	43.1	42.2	41.2	43.1	42.2	42.3	44.4	43.3
Log (short pause duration)	51.3	50.2	50.7	50.6	52.0	51.4	49.8	51.4	50.6
Log (medium pause duration)	40.2	42.0	41.3	40.9	42.3	41.6	40.5	41.3	41.0
Log (long pause duration)	46.0	44.3	45.2	46.5	44.9	45.8	41.0	39.3	40.1

Table 3.2: EER (%) for each prosodic feature using different distance measurements and k=3.

The best individual results are achieved in the features related to fundamental frequency, specially its mean and maximum values. On the contrary, the worse results are obtained when using features related to pauses. This could be explained by the fact that, since we are dealing with speech conversations between pairs of speakers, pauses depend somehow on the speech interruptions of the other speaker, in both frequency and duration terms. And since each conversation is carried out with different pairs of speakers, the way in which the speaker is interrupted varies every time.

Regarding the distances used in the classification step, the best results are clearly achieved with the Kullback-Leibler divergence, while Euclidean and Mahalanobis distances perform in a similar way. On the other hand, no considerable differences seem to appear between the use of the first and the third nearest neighbour classifier. Nevertheless, by applying k=3 the results are slightly better, specially when Euclidean and Mahalanobis distances are used.

In Figure 3.2 the EER of every prosodic feature using each one of the three distances is plotted for comparison. The results correspond to the use of splits 1-6 of the Switchboard-I database (splits 1-3 for training and 4-6 as cohort speakers and viceversa) and the 3rd nearest neighbour for classification. The graph shows that —except for the features related to pauses and the F0 pseudo-slope— Kullback-Leilber divergence gives, in this case, much better results than the other distances.

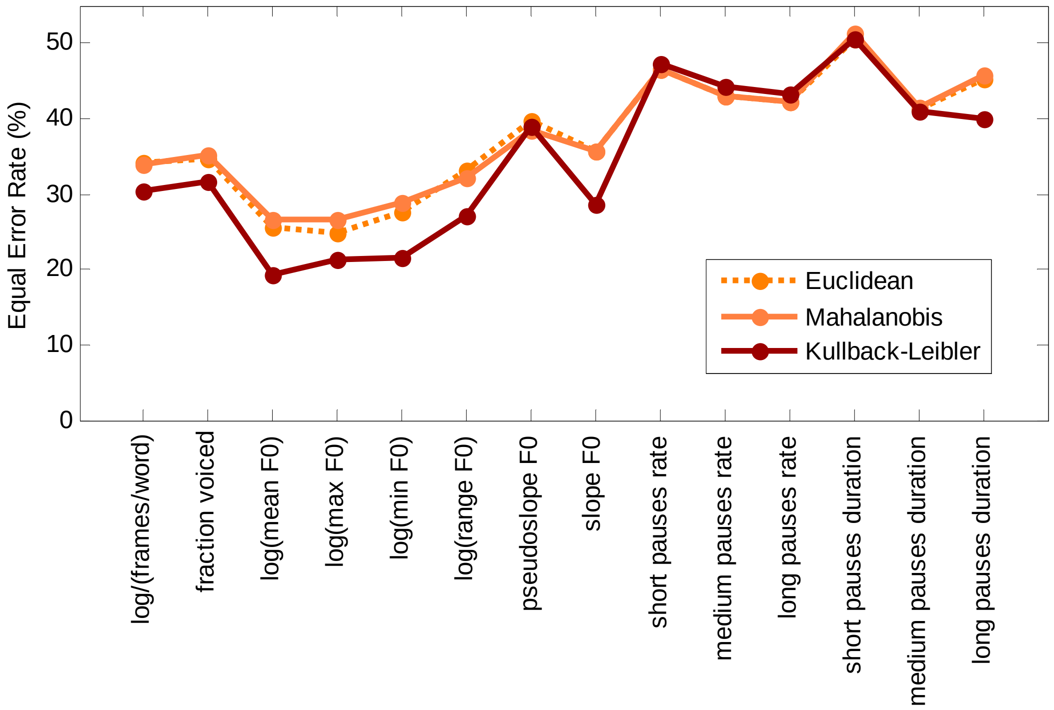


Figure 3.2: EER for each prosodic feature using three different distances: Euclidean, Mahalanobis and Kullback-Leibler divergence, the 3rd nearest neighbour and splits 1-6.

3.2.3 Fusion of prosodic features

This section contains the results obtained after fusing the individual prosodic parameters. Fusion of prosodic features has been performed at the score level, using two linear fusion techniques —**simple sum** and **matcher weighting**— combined with two conventional normalisation methods: **min-max** and **z-score** (see section 2.4.2).

Splits 1-3 of the Switchboard-I database have been used to train the speaker models, using splits 4-6 as cohort speakers. The complementary set (splits 4-6 for training and splits 1-3 as cohort speakers) is used as a developing set in order to obtain the weights for matcher weighting fusion and the statistical values needed for the normalisation phase (minimum, maximum, mean and standard deviation).

The fusion results are shown in Table 3.3. The first three rows contain the results after fusing each of the three prosodic feature related sets: segment duration, fundamental frequency and pauses. Note that only five pauses are used in the pauses set instead of six (as originally). This is due to the fact that the *logarithm of the short pause duration* feature was removed from the list, since most of its individual EER values were above 50% (see Tables 3.1 and 3.2).

Feature set	**Simple sum**		**Matcher weighting**	
	min-max	z-score	min-max	z-score
2 segment duration	26.0	24.9	25.6	**24.8**
6 fundamental frequency	18.6	18.7	18.7	**18.3**
5 pauses	38.5	38.4	38.4	**38.2**
2 segment duration + 6 F0	15.0	15.6	**14.9**	15.3
2 segment duration + 5 pauses	31.8	30.6	29.9	**29.3**
6 F0 + 5 pauses	23.5	25.5	**21.0**	23.1
all features	21.6	22.4	19.2	**19.0**

Table 3.3: EER (%) for each prosodic feature combination using min-max and z-score normalisations, and simple sum and matcher weighting fusion techniques.

The lowest equal error rates are achieved, once again, by using those features related to fundamental frequency, while the worse results are given by fusing pause-related features. In the first three rows —where fusion is carried out within a specific set of features— the best results are obtained by z-score normalisation plus matcher weighting technique. When using all 13 prosodic features, the best result is also achieved with the z-score and matcher weighting combination. Unlike the previous cases, in this overall fusion the simple sum technique is clearly outperformed by matcher weighting fusion. This is also shown in Figure 3.3, where the DET curves corresponding to the fusion of all the prosodic parameters are plotted. Although in matcher weighting fusion the EER point is lower when normalising with z-score, the DET curves show that the overall performance is better when using min-max normalisation.

The performance of the sets related to segment duration and F0 features is clearly improved when both sets are combined. This improvement can also be seen in the corresponding DET curves plotted in Figure 3.4, using z-score normalisation and matcher weighting fusion. Nevertheless, pause information is not useful enough to improve neither the segment duration nor the F0 related features, so that the best performance of the system is achieved when pause features are not considered.

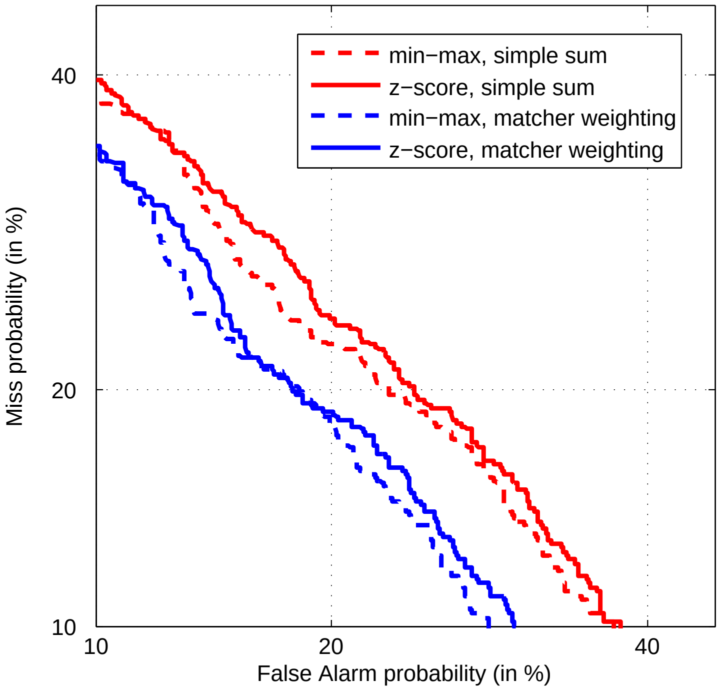


Figure 3.3: DET curves for the overall fusion of 13 prosodic features combining two normalisation techniques (min-max and z-score) with two types of fusion (simple sum and matcher weighting).

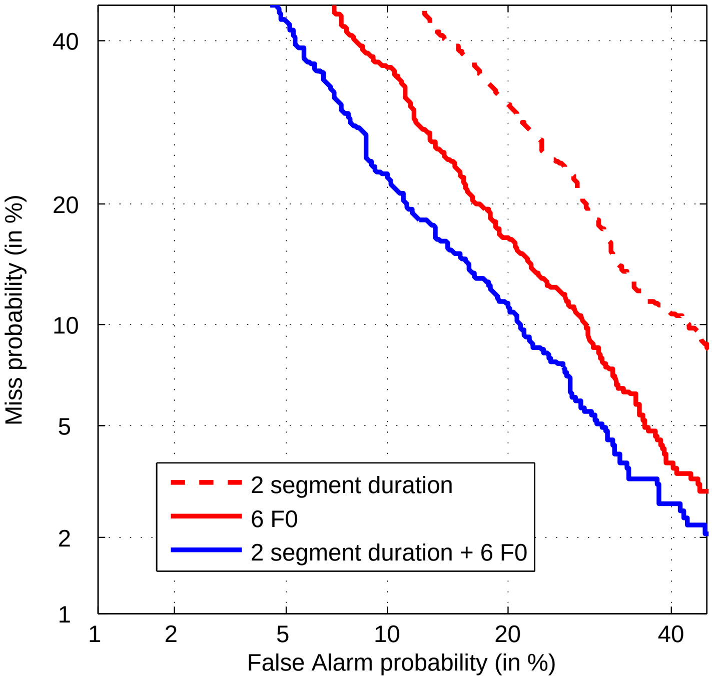


Figure 3.4: DET curve showing the fusion of 2 features related to segment duration, the fusion of 6 features related to fundamental frequency, and fusion of all 8 features, using z-score normalisation and matcher weighting technique.

3.3 Fusion of spectral and prosodic parameters

In this section, the above-proposed prosodic system is fused with the spectral information. As in the previous section, the speaker models of both prosodic and spectral systems were trained with 8 conversation sides, using splits 1-3 of Switchboard-I database as a training set. The systems were tested with one-conversation-side according to the NIST's 2001 Extended Data Task.

Fusion was also performed at the score level for splits 1-3, using simple sum and matcher weighting techniques combined with two previous conventional normalisations: min-max and z-score. On the other hand, splits 4-6 were used to train the weights and the normalisation statistical values.

Tables 3.4 and 3.5 show that all sets of prosodic parameters are clearly improved when adding spectral information. Nevertheless, the threshold of EER=10.1% corresponding to the performance of the spectral system is not always outperformed when both prosodic and spectral systems are combined. When fusion is carried out by means of simple sum technique, only EER values below 10.1% are achieved where the F0-based set is involved. However, although the performance of the combined *segment duration* and F0 sets is higher than when used alone, this better performance is not always maintained when spectral parameters are added: in Table 3.4, the best verification results are achieved by fusing only spectral and fundamental frequency related parameters.

Feature set	**Min-max**		**Z-score**	
	prosody	+ spectrum	prosody	+ spectrum
2 segment duration	26.0	10.4	24.9	10.7
6 fundamental frequency	18.6	**8.9**	18.7	**8.8**
5 pauses	38.5	21.3	38.4	24.6
2 segment duration + 6 F0	15.0	**9.2**	15.6	**9.4**
2 segment duration + 5 pauses	31.8	17.7	30.6	21.0
6 F0 + 5 pauses	23.5	15.0	25.5	18.2
all features	21.6	13.8	22.4	15.6

Table 3.4: EER (%) for prosodic feature sets combined with spectral parameters, using min-max and z-score normalisations and simple sum fusion.

Feature set	**Min-max**		**Z-score**	
	prosody	+ spectrum	prosody	+ spectrum
2 segment duration	25.6	**9.2**	24.8	**9.2**
6 fundamental frequency	18.7	**8.5**	18.3	**8.2**
5 pauses	38.4	11.7	38.2	13.7
2 segment duration + 6 F0	14.9	**7.7**	15.3	**7.7**
2 segment duration + 5 pauses	29.9	11.0	29.3	12.6
6 F0 + 5 pauses	21.0	**9.4**	23.1	10.6
all features	19.2	**8.1**	19.0	**9.5**

Table 3.5: EER (%) for prosodic feature sets combined with spectral parameters, using min-max and z-score normalisations and matcher weighting fusion.

On the other hand, Table 3.5 shows that when fusion is performed by means of matcher weighting technique, the overall performance in these experiments is clearly improved. The threshold of EER=10.1% corresponding to the spectral system is outperformed when both *segment duration* and F0-related sets are involved alone or in combination. Moreover, the performance is also improved by combining the spectral information with the F0- and pause-related sets using a min-max normalisation before matcher weighting fusion. The difference of system performance between both types of fusion is much larger here than when the prosodic features were used alone. This could be due to the fact that the EER of the spectral parameters is considerably lower than the EER of all the prosodic parameters; hence the spectral weight is larger and the influence of the spectral system —which has the best performance— is bigger, leading to better overall results.

There are in general no big differences between using min-max and z-score normalisation with other parameters, but when pauses are involved in a z-score normalisation, both alone or in combination with other sets of features, the results are considerably worse than when using a min-max normalisation instead (see Tables 3.4 and 3.5). Since z-score is optimal for Gaussian data, this could suggest that the score distribution of the pause-related parameters differs to a large extent from a Gaussian distribution.

3.4 Jitter and shimmer

Fundamental frequency is determined physiologically by the number of cycles that the vocal folds do in a second. Jitter refers to the variability of fundamental frequency, and it is affected mainly due to the lack of control of vocal fold vibration (Behlau et al., 2001; Wertzner et al., 2005). On the other hand, vocal intensity is related to subglottic pressure of the air column, which, in turn, depends on other factors such as amplitude of vibration and tension of vocal folds (Behlau and Pontes, 1995). Shimmer is affected mainly because of the reduction this tension and mass lesions in the vocal folds (Behlau et al., 2001).

Both jitter and shimmer features have been largely used to detect voice pathologies (see, e.g. Wagner (1995); Michaelis et al. (1998); Kreiman and Gerrat (2005)). They are commonly measured for long sustained vowels, and values of jitter and shimmer above a certain threshold are considered being related to pathological voices, which are usually perceived by humans as breathy, rough or hoarse voices. Since pathological voices normally characterise a particular speaker, this leads to think that both jitter and shimmer features might be useful to distinguish speakers.

Some recent works and current literature (Linville, 2001; Schötz, 2001; Minematsu et al., 2002; Wittig and Müller, 2005) have shown that jitter and shimmer are appropriate features to characterise speakers —specifically the age and the gender of the individual. On the other hand, Slyh et al. (1999) and Li et al. (2005) reported that significant differences can occur in jitter and shimmer measurements between different speaking styles, especially in shimmer measurement. Nevertheless, prosody is also highly-dependent on the emotion of the speaker, and prosodic features are useful in automatic recognition systems even when no emotional state is distinguished, which leads to the hypothesis that jitter and shimmer features can be also useful in the speaker recognition task.

3.4.1 Jitter and shimmer measurements

The novel component in this section is the analysis of jitter and shimmer features in order to test their usefulness in speaker verification. These features have been extracted by using the Praat voice analysis software (Boersma and Weenink, 1992). Praat reports different kinds of measurements for both jitter and shimmer features, which are listed below.

Jitter measurements

Jitter (absolute) is the cycle-to-cycle variation of fundamental frequency, i.e. the average absolute difference between consecutive periods, expressed as:

$$\text{Jitter (absolute)} = \frac{1}{N}\sum_{i=1}^{N-1} |T_i - T_{i+1}| \,, \tag{3.5}$$

where T_i are the extracted F0 period lengths and N is the number of extracted F0 periods, as shown in Figure 3.5. The Multidimensional Voice Program (MDVP) from the Kay Elemetrics Corporation (Deliyski, 1993) calls this parameter *Jita*, and gives 83.200 μs as a threshold for pathology.

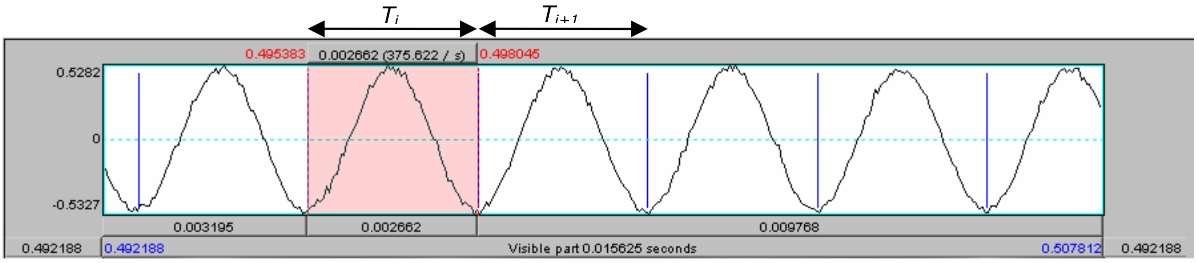


Figure 3.5: Jitter measurement for N=5 F0 periods.

Jitter (relative) is the average absolute difference between consecutive periods, divided by the average period. It is expressed as a percentage:

$$\text{Jitter (relative)} = \frac{\frac{1}{N}\sum_{i=1}^{N-1} |T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \,. \tag{3.6}$$

MDVP calls this parameter *Jitt*, and gives 1.040% as a threshold for pathology.

Jitter (rap) is defined as the Relative Average Perturbation, the average absolute difference between a period and average of it and its two neighbours, divided by the average period. MDVP gives 0.680% as a threshold for pathology.

Jitter (ppq5) is the five-point Period Perturbation Quotient, computed as the average absolute difference between a period and the average of it and its four closest neighbours, divided by the average period. MDVP calls this parameter PPQ, and gives 0.840% as a threshold for pathology.

Shimmer measurements

Shimmer (absolute) is expressed as the variability of the peak-to-peak amplitude in decibels, i.e. the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20:

$$\text{Shimmer (absolute)} = \frac{1}{N-1}\sum_{i=1}^{N-1} \left|20\log(A_{i+1}/A_i)\right| \,, \tag{3.7}$$

where A_i are the extracted peak-to-peak amplitude data and N is the number of extracted fundamental frequency periods, as shown in Figure 3.6. MDVP calls this parameter *ShdB*, and gives 0.350 dB as a threshold for pathology.

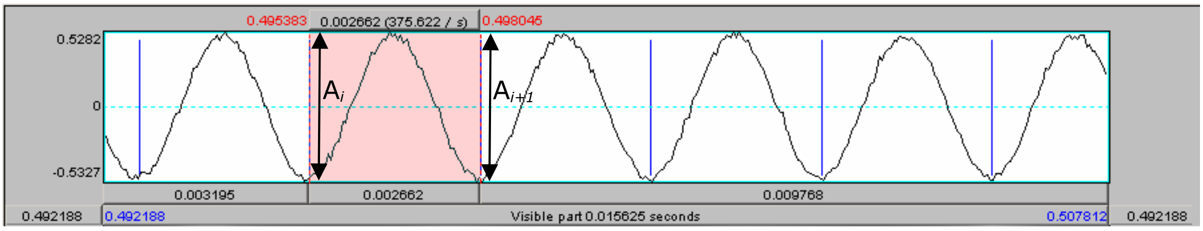


Figure 3.6: Shimmer measurement for N=5 F0 periods.

Shimmer (relative) is defined as the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude, expressed as a percentage:

$$\text{Shimmer (relative)} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \,. \tag{3.8}$$

MDVP calls this parameter *Shim*, and gives 3.810% as a threshold for pathology.

Shimmer (apq3) is the three-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbours, divided by the average amplitude.

Shimmer (apq5) is defined as the five-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its four closest neighbours, divided by the average amplitude.

Shimmer (apq11) is expressed as the 11-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its ten closest neighbours, divided by the average amplitude. MDVP calls this parameter APQ, and gives 3.070% as a threshold for pathology.

3.4.2 Verification experiments

Experimental setup

All the recognition experiments described in this section have been performed over the Switchboard-I database. A nine-feature vector was extracted for the acoustic system based on the nine jitter and shimmer measurements described above. As in the F0-related features of the prosodic system, features were extracted using the Praat software for acoustic analysis (Boersma and Weenink, 1992), performing an acoustic periodicity detection based on a cross-correlation method, with a window length of 40/3 ms and a shift of 10/3 ms. Mean over all words and standard deviation were computed for each individual measurement.

The systems used 8 conversation sides to train the speaker models. Training was performed using splits 1-3 of Switchboard-I database and splits 4-6 as cohort speakers, according to the NIST's 2001 Extended Data task. Testing was performed by computing the distance between the test feature vector and the k feature vectors of the claimed speaker, using the k-nearest neighbour method with k=3 and the symmetrised Kullback-Leibler divergence. Fusion of individual features was performed at the score level for splits 1-3, using the matcher weighting method and z-score normalization. Weights were trained from the splits 4-6 using splits 1-3 as cohort speakers.

Results

First of all, the EER for each individual jitter and shimmer features is computed. Table 3.6 and Table 3.7 show the EER results for jitter and shimmer measurements, respectively. Both tables give the EER for the individual measurements and the combination of the measurement set.

Jitter measurement	**EER (%)**
Jitter (absolute)	26.9
Jitter (relative)	33.7
Jitter (rap)	34.2
Jitter (ppq5)	33.8
Fusion	**29.2**

Table 3.6: EER for jitter measurements.

Shimmer measurement	**EER (%)**
Shimmer (absolute)	26.9
Shimmer (relative)	28.9
Shimmer (apq3)	28.1
Shimmer (apq5)	32.9
Shimmer (apq11)	33.8
Fusion	**25.5**

Table 3.7: EER for shimmer measurements.

The results show that at least both absolute measurements of jitter and shimmer are potentially useful in speaker recognition. In the case of jitter, its relative measurements do not seem to supply helpful information, since the fusion of all jitter measurements does not outperform the result obtained with the isolated absolute measurement. In order to ensure this assumption, the absolute measurement of jitter was fused with the best-performing relative measurement: the *Jitter (relative)*. The combination of both measurements provided an EER of 29.3%, so that fusion of both measurements does not improve the absolute jitter measurement either.

In the case of shimmer measurements, their final fusion improves slightly the best isolated results *(Shimmer (absolute))*. Since all relative measurements of the same feature are highly correlated, only the relative measurement of shimmer giving the best EER is used: the *Shimmer (apq3)*. To ensure that this measurement provides complementary information to *Shimmer (absolute)*, both measurement were combined. The EER obtained in the fusion equaled 26.3%, improving slightly the isolated absolute measurement of shimmer.

From now on, only three cycle-to-cycle variability measurements will be used as new features:

- the absolute measurement of *jitter* (average absolute difference between consecutive periods)
- the absolute measurement of *shimmer* (average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods)
- one of the relative measurements of *shimmer*: the three point Amplitude Perturbation Quotient.

This set of three measurements will be referred to as the *JitShim* system. As shown in Table 3.8, the EER of the combination of these measurements equaled 22.5%.

JitShim system	**EER (%)**
Jitter (absolute)	26.9
Shimmer (absolute)	26.9
Shimmer (apq3)	28.1
Fusion	**22.5**

Table 3.8: EER of jitter and shimmer measurements used in the *JitShim* system and EER of their combination.

3.5 Fusion of jitter and shimmer with prosodic and spectral features

In order to see how jitter and shimmer are able to improve the prosodic and the voice spectral based recognition systems, the new features are added to both systems separately, which are the ones introduced in sections 3.1 and 3.2.

All three systems used 8 conversation sides to train the speaker models. Training was performed using splits 1-3 of Switchboard-I database using splits 4-6 as cohort speakers, according to the NIST's 2001 Extended Data task. Fusion of individual features was performed at the score level for splits 1-3, using the matcher weighting method and z-score normalization. Weights were trained from the splits 4-6 using splits 1-3 as cohort speakers.

First of all, all eight prosodic features used in the baseline system are combined with the three features of the novel *JitShim* system, resulting in a new eleven-featured system. Second, the *JitShim* system is added to the voice spectral baseline system. This allows comparing how complementary jitter and shimmer are to prosodic and spectral features, respectively. Finally, the *JitShim* system is combined with both baselines, in order to see how the new features improve the speaker verification system. The results of these experiments are shown in Table 3.9 and their DET curves are plotted in Figure 3.7. The EER before the introduction of the *JitShim* system are given in the middle column of the table, and results after adding jitter and shimmer features are shown in the right column.

Baseline system		+ JitShim
Prosodic	15.3 ± 1.6	13.1 ± 1.5
Spectral	10.1 ± 1.4	8.6 ± 1.3
Fusion	**7.7** ± 1.2	**6.8** ± 1.1

Table 3.9: EER (%) for prosodic and spectral systems before and after adding jitter and shimmer features.

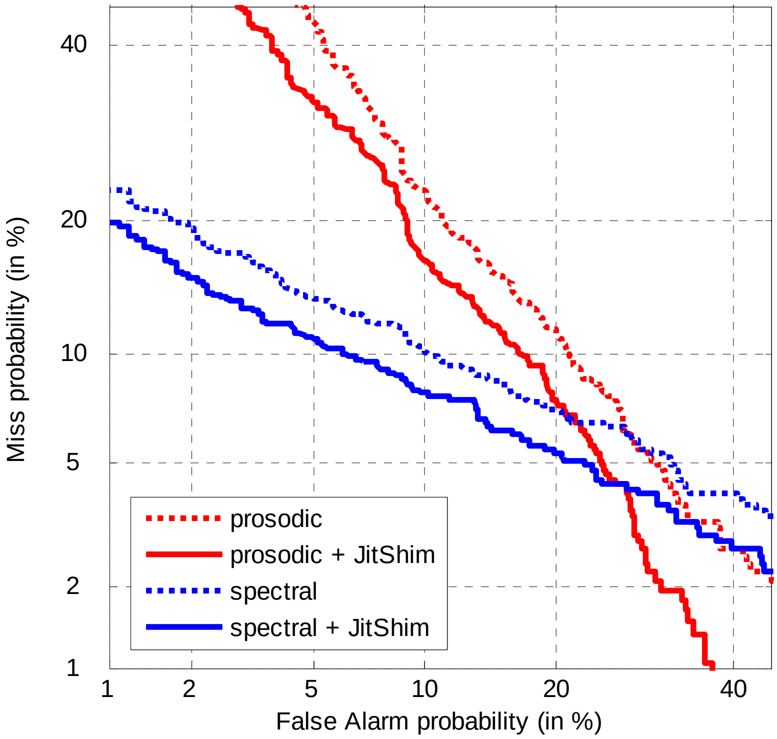


Figure 3.7: DET curves for prosodic and spectral systems before and after adding jitter and shimmer features.

The results and the DET curves plotted in Figure 3.7 show that both prosodic and spectral baselines are clearly improved when jitter and shimmer features are added to the systems. The best relative improvement is achieved by adding the *JitShim* system to the spectral system (15%), although when fusing *JitShim* with the prosody based system the improvement is also considerable (14%). That suggests that the information provided by jitter and shimmer measurements to prosodic parameters and the information supplied to the spectral system are, in this case, equally complementary.

The speaker verification system based on prosodic and spectral parameters is also improved by adding the *JitShim* system, as it can be seen in the DET curves plotted in Figure 3.8, achieving the lowest EER equaling 6.8%. So, jitter and shimmer features seem to be useful in speaker recognition and should be considered in future experiments.

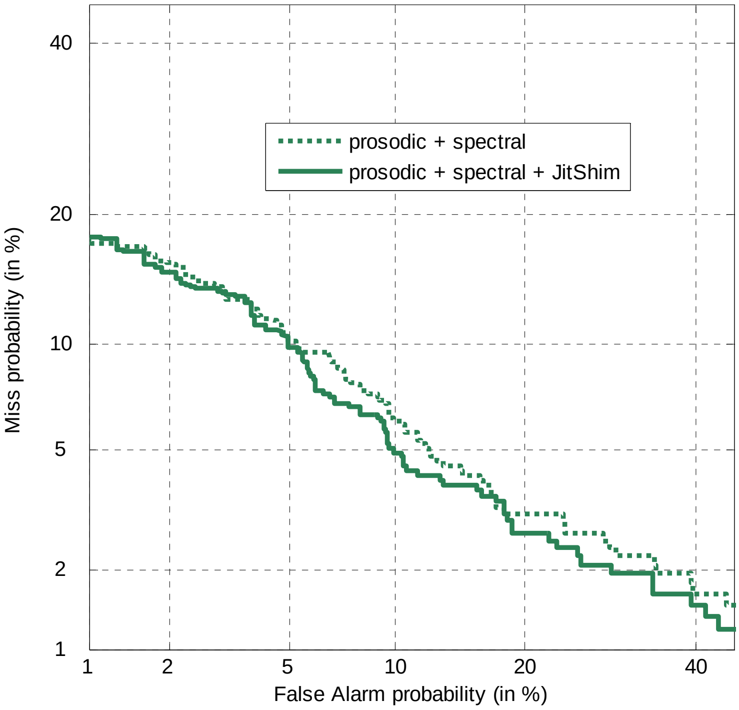


Figure 3.8: DET plot showing the improvement of the baseline system after adding jitter and shimmer.

3.6 Conclusions

Several works have demonstrated that the use of prosodic information helps to improve recognition systems based solely on spectral —filter— parameters. The experiments performed in this section corroborate this fact. A preliminary speaker verification system based on prosodic features has been built in order to improve a voice spectrum based verification system over the conversational Switchboard-I database. Whereas the performance of a spectral system in these experiments is considerably improved by using those prosodic features related to segment duration and fundamental frequency, the information contained in the pauses was not useful to improve either of the spectral and the rest of prosodic features.

Although prosodic information seems to be normally helpful, the way how the different scores are fused and the distance used in the decision classifier are important. Kullback-Leibler divergence in the nearest neighbour classifier outperforms —in almost all cases— Euclidean and Mahalanobis distances, as seen in section 3.2.2. Moreover, no big differences seem to appear between the choice of min-max and z-score normalisations, except when pauses and spectral information are involved, where the performance by using z-score gets worse in comparison with the use of min-max normalisation. Furthermore, the use of matcher weighting technique outperforms, in most cases, the use of simple sum technique, as shown in sections 3.2.3 and 3.3.

In section 3.5, a speaker verification system based on prosodic and spectral parameters has been improved by adding jitter and shimmer features, which analyse the perturbation of fundamental frequency and waveform amplitude, respectively. In these experiments, the absolute measurements of both features seem to be more discriminant than their relative measurements. Furthermore, the results show that jitter and shimmer can provide complementary information to both spectral and prosodic systems.

Chapter 4

A Voice and Face Biometric System using HE-SVM

As it was seen in section 2.1, a multimodal biometric system involves the combination of two or more human biometric characteristics in order to achieve better results than using unimodal recognition (Bolle et al., 2004). In this chapter, two different modalities —voice and face— will be fused together in order to achieve better results. As a unimodal biometric characteristic, voice can be represented in several levels of information (see section 2.2.2). In the experiments performed in this chapter, two different levels of voice information will be used: spectral and prosodic, and although both levels are related to speech, they will be referred to as different modalities.

The speech and face parameters used in multimodal biometric recognition are described in section 4.1. Histogram equalisation as a score normalisation technique is presented in section 4.2. The description of the multimodal database is included in 4.3, and section 4.4 deals with the combination of voice and face information to obtain a multimodal biometric system. This includes the use of several strategies based on the fusion of scores in different steps (4.4.1), and the score normalisation by means of histogram equalisation in order to improve the overall biometric fusion (4.4.2). Since the computation cost with the preliminary biometric database was very large, further experiments were performed by using a reduced database (4.4.3). Finally, conclusions of the experiments are presented in section 4.5.

4.1 Speech and face parameters

Multimodal person recognition systems normally use short-term spectral features as voice information. However, it has been shown that prosody helps to improve voice spectrum based speaker recognition systems. Therefore, prosodic features may be used in multimodal person verification in order to achieve better results.

In this chapter, prosodic and vocal tract spectral features are used in combination with facial parameters. The speech parameters, which are listed in Table 4.1, are the same as used in chapter 3. Due to their low performance, the prosodic parameters related to pauses are not considered.

Speech parameters	
Prosodic	Log (frames/word)
	Fraction of voiced segments
	Fraction of unvoiced segments
	Log (mean F0)
	Log (max F0)
	Log (min F0)
	Log (range F0)
	F0 slope
	Stylised F0 slope
Spectral	Frequency Filtering + Δ and $\Delta\Delta$ coefficients

Table 4.1: Speech parameters used in the multimodal biometric system.

Note that the parameters *fraction of voiced segments* and *fraction of unvoiced segments* are completely correlated. In chapter 3, both parameters were initially used, and the problem was later solved by removing the second parameter and recalculating the results. However, the fusion experiments performed in this chapter have a high computational cost. In order to solve the problem, the results were recalculated over a reduced multimodal database (see section 4.4.3).

Facial recognition systems are based on the conceptualisation that a face can be represented as a collection of sparsely distributed parts: eyes, nose, cheeks, mouth, etc. Non-negative matrix factorization (NMF), introduced in Lee and Seung (2001), is used in this work to yield a sparse representation of localised features in order to represent the constituent facial parts.

NMF is an appearance-based face recognition technique based on the conventional component analysis techniques. Given a facial image database denoted by a matrix X, NMF aims to factorise the matrix X into (usually) two matrices Z and H; i.e. to find two matrices Z and H such that $X \approx ZH$. The facial image u_j after the NMF decomposition can be written as $u_j \approx Zh_j$, where h_j is the j-th column of H. Thus, the rows of the matrix Z can be considered as basis images and the vector h_j as the corresponding weight vector. The h_j vectors can also be considered as the projected vectors of a lower dimensional feature space. Since $x_j \approx Zh_j$, a natural way to compute the projection of x_j to a lower dimensional feature space using NMF is $x'_j = (Z^T Z)^{-1} Z^T x_j$.

Factorisation of matrices is generally non-unique; therefore, different matrix factorisation methods have been developed (e.g. principal component analysis, singular value decomposition, etc.) by incorporating different constraints. Non-negative matrix factorisation imposes non-negative constraints in both elements of the matrices Z and H; thus, only additive —non-subtractive— combinations are allowed. This is believed to correspond better to the intuitive notion of combining facial parts in order to create a complete face.

One of the algorithms initially proposed for finding the matrices Z and H used the following metric:

$$D_N(X \parallel ZH) = \sum_{i,j} \left(x_{i,j} \ln \left(\frac{x_{i,j}}{\sum_l z_{i,l} h_{l,j}} \right) + \sum_k z_{i,k} h_{k,j} - x_{i,j} \right) \tag{4.1}$$

as the measure of the cost for factorising X into ZH (Lee and Seung, 2001).

The NMF factorisation is the outcome of optimisation: $\min_{Z,H} D_N(X\|ZH)$ subject to the non-negative constraints of the matrices Z and H:

$$z_{i,k} \geq 0, \qquad h_{k,j} \geq 0, \qquad \sum_i z_{i,j} = 1, \qquad \forall j \tag{4.2}$$

By using an auxiliary function and the Expectation Maximisation algorithm (Lee and Seung, 2001), specific update rules for both Z and H guarantee a non-increasing behaviour of 4.1.

NMF does not use the information about how the various facial images are separated into different facial classes. The most straightforward way to exploit discriminative information in NMF is to try to discover discriminative projections for the facial image vectors after the projection to the basis image matrix.

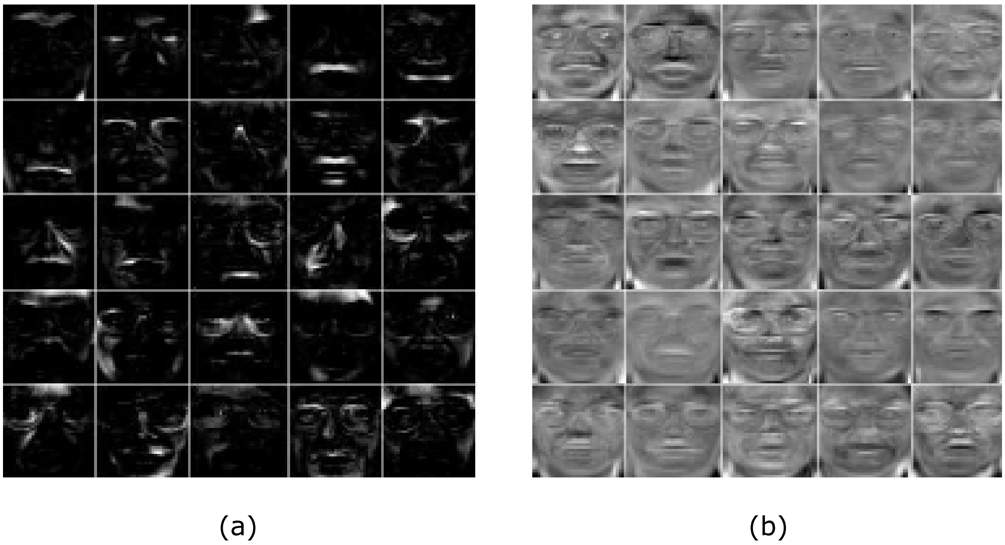


Figure 4.1: A set of 25 basis images for (a) NMF and (b) NMF-faces.

There is no upper limit for how many bases someone can construct using NMF decomposition in the update rule for the matrix Z, and unless a limited number of bases by NMF is created, the within scatter matrix for the coefficient vectors h_j is singular (i.e. the matrix is not invertible). In order to solve this problem, the Fisherfaces approach (Belhumeur et al., 1997) is used. The face recognition scores used in these experiments have been calculated in this way with the NMF-faces method (Zafeiriou et al., 2005a,b), in which the final basis images are closer to facial parts. A set of several basis images for both NMF and NMF-faces methods is illustrated in Figure 4.1.

4.2 Histogram equalisation

Histogram equalisation (HE) is a non-linear transformation that converts a probability distribution to another. The aim of this transformation is to match the statistical properties (mean, variance, skew, kurtosis, etc.) of two probability distributions.

Histogram equalisation is a widely used non-linear method designed for the enhancement of images. It employs a monotonic, non-linear mapping that reassigns the intensity values of pixels in the input image. The aim is to control the shape of the output image intensity histogram in order to achieve a uniform distribution of intensities or to highlight certain intensity levels.

This technique has been also developed for speech recognition adaptation approaches and correction of non-linear effects typically introduced by speech systems such as microphones, amplifiers, clipping and boosting circuits and automatic gain control circuits (Balchandran and Mammone, 1998; Hilger and Ney, 2001).

The objective of HE is to find a non-linear transformation that aims to reduce the mismatch of the statistics of two signals. In Pelecanos and Sridharan (2001); Skosan and Mashao (2006) this concept was applied to the acoustic features in order to improve the robustness of a speaker verification system by reducing the mismatch between training and test conditions and the additive noise and channel and transducer effects.

In this thesis, histogram equalisation is applied to the score distributions. To this end, a matching of the cumulative distribution function (CDF) of a reference distribution and the CDF of the variable to be transformed is performed as follows (de la Torre et al., 2005; Skosan and Mashao, 2006):

Let x be a random variable with a probability distribution $p_x(x)$, and let $y = T(x)$ be a single-valued and monotonically increasing transformation function that converts the probability distribution $p_x(x)$ into a reference probability distribution $p_{ref}(y)$. The transformation $T(x)$ then makes the probability of finding x in the differential range dx equal to the probability of finding y in the differential range dy, i.e.:

$$p_{ref}(y)dy = p_x(x)dx \ , \tag{4.3}$$

and modifies the original probability distribution $p_x(x)$ according to the expression

$$p_{ref}(y) = p_x(x)\frac{dx}{dy} = p(G(y))\frac{dG(y)}{dy} \ , \tag{4.4}$$

where $G(y) = x$ is the inverse of $T(x)$. Using Equation 4.4, the cumulative distribution functions associated with $p_x(x)$ and $p_{ref}(y)$ are related as follows:

$$\begin{aligned} C_x(x) &= \int_{-\infty}^{x} p_x(x')dx' = \int_{-\infty}^{T(x)} p_x(G(y)')\frac{dG(y)}{dy}dy' \\ &= \int_{-\infty}^{y} p_{ref}(y')dy' = C_{ref}(y) = C_{ref}(T(x)). \end{aligned} \tag{4.5}$$

Thus, the transformation $T(x)$ is given by:

$$T(x) = C_{ref}^{-1}(C_x(x)) \ , \tag{4.6}$$

where C_{ref}^{-1} is the inverse of the CDF of the reference probability distribution.

Nevertheless, only a finite number of observations are usually available in practical implementations, and cumulative histograms are used instead of cumulative probabilities; for this reason, the transformation is called *histogram equalisation*. A schematic diagram in Figure 4.2 shows the matching process of the cumulative histogram of the original variable x and the reference cumulative histogram.

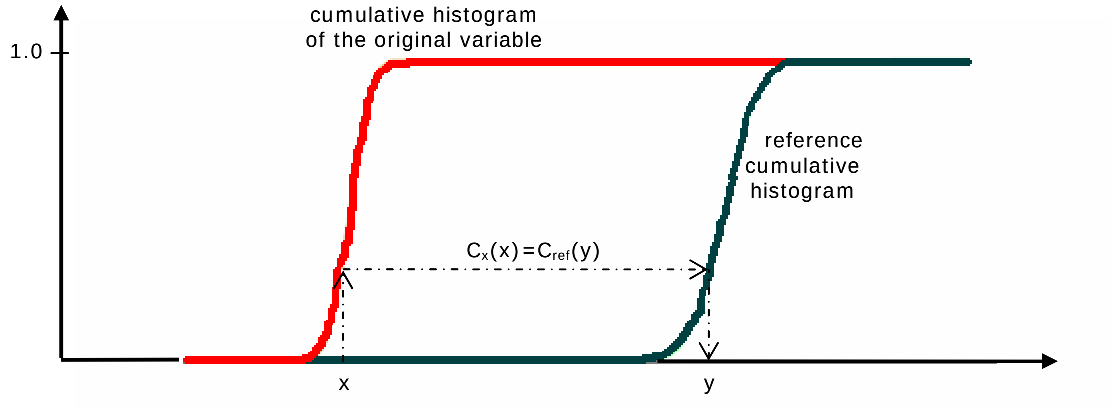


Figure 4.2: Histogram equalisation: matching of the cumulative distribution.

Practical implementation

The aim of histogram equalisation is to transform the score distributions obtained for each modality, so that their statistics are equal to those of a reference distribution. Therefore, the first step is to select a suitable reference distribution $p_{ref}(y)$ corresponding to one of the modalities and compute its cumulative histogram $C_{ref}(y)$. Next, HE can be applied to the score distribution of each modality as follows:

1. Determine the maximum (x_{max}) and minimum (x_{min}) values across the whole set of observations —scores, in this case— of a particular modality.

2. Divide the range $[x_{min}, x_{max}]$ into M equally-spaced non-overlapping intervals B_i satisfying the following conditions:

$$\begin{aligned} x_{min} &= b_1 < b_2 < ... < b_{M+1} \\ B_i &= [b_i, b_{i+1}) \end{aligned} \tag{4.7}$$

3. Construct a histogram of the scores in the set using the intervals B_i; i.e., scan the set and count the number of scores (observations) falling into each interval B_i.

4. Compute the normalised version of the histogram obtained by using the following expression:

$$p_x(x \in B_i) = \frac{n_i}{N_x} \tag{4.8}$$

being n_i the number of observations in the interval B_i, and N_x the total number of observations in the set. In fact, Equation 4.8 is an approximation of the probability of x lying in the interval B_i.

5. Compute the cumulative histogram of the set, using the normalised histogram constructed in the previous step, as follows:

$$C_x(x : x \in B_i) = \sum_{j=1}^{i} \frac{n_j}{N_x} \tag{4.9}$$

which is an approximation of the true cumulative distribution function.

6. Replace each value of x by the value of y that corresponds to the same point in the reference and computed cumulative histograms, so that $C_x(x) = C_{ref}(y)$.

4.3 Multimodal database

Recognition experiments have been performed over the Switchboard-I speech database and the video and speech XM2VTS database of the University of Surrey (Lüttin and Maître, 1998). Switchboard-I database has been used for the speaker recognition experiments. Speaker scores have been obtained by using two different systems: a voice spectrum based recognition system and a prosody based recognition system (see chapter 3), whose features are also listed in section 4.1.

Face recognition experiments have been performed over the XM2VTS database, which is a multimodal database consisting of face images, video sequences and speech recordings of 295 subjects. Only the face images (four frontal face images per subject) have been used in these experiments. In order to evaluate verification algorithms on the database, the evaluation protocol described in Lüttin and Maître (1998) was followed. The well-known Fisher discriminant criterion was constructed as Belhumeur et al. (1997) in order to discover discriminant linear projections and to obtain the facial scores.

In the fusion step, the scores obtained from the speech recognition experiments have been combined with the scores obtained from the face recognition experiments. Since both databases contain biometric characteristics belonging to different users, a chimerical database has been built to perform the experiments. A chimerical database is an artificial database created by linking two or more unimodal biometric characteristics from different individuals in order to form artificial —or chimerical— users.

Due to the large number of experiments needed for a statistically adequate number of errors, it was necessary to relate one user from one database to more than one user from the other database. In these experiments, the chimerical database consists of 20661 users created by combining 179 users of the splits 1-3 of the Switchboard-I database and 270 users of the XM2VTS database. The scores were then split into two equal sets (development and test) for each recognition system, obtaining a total amount of 46500 scores for each set: 16800 client trials and 29700 impostor trials.

4.4 Multimodal fusion experiments

As it was seen in section 2.4, different fusion and normalization techniques can be used in order to fuse several unimodal biometrics. The results obtained can rather vary depending on the methods used in the fusion. In this section, several fusion strategies are proposed in order to fuse the three different modalities. These strategies are based on the fusion of unimodal scores in one, two or three steps (Farrús et al., 2006b).

Apart from the above-mentioned fusion strategies, several normalisation techniques are applied before the score fusion, namely, the conventional linear z-score normalisation and the non-linear technique histogram equalisation (see section 4.2).

Both speech modalities based on prosodic and spectral features are taken from the Switchboard-I database. In order to get a multimodal database including both facial and conversational speech features, an artificial database is created combining the faces of the XM2VTS with the voices of Switchboard-I. The first part of the experiments are performed with an extensive multimodal database. Later, a smaller database is used in order to reduce the computational cost of the experiments.

4.4.1 Fusion strategies using MW and SVM

Matcher weighting (MW) fusion is one of the most conventional fusion techniques, where each unimodal score is weighted by a factor proportional to each recognition rate (see section 2.4.2). In contrast to MW, non-linear and machine learning based techniques —such as support vector machines— may lead to a higher performance.

Learning based fusion can be treated as a pattern classification problem in which the scores obtained by individual classifiers are seen as input patterns to be labeled as *accepted* or *rejected*. A support vector machine (SVM) is a binary classifer that —in order to separate the data— uses a non-linear decision function, achieving an extension to non-linear boundaries by using a specific kernel function (see 2.4.2).

In this section, the unimodal scores are fused by using both MW and SVM techniques. The kernel function used in SVM in these experiments is a radial basis function (RBF) expressed as:

$$K(x_i, x_j) = \exp\left[-\frac{1}{2}\left(\frac{\|x_i - x_j\|}{\sigma}\right)^2 \right] \tag{4.10}$$

and the constant C (Eq. 2.30) is set to $C = 100$.

A conventional z-score normalisation will be applied before MW fusion technique. Support vector machines seem to be used alone without any prior normalisation; however, an intrinsec min-max normalisation is, in these cases, normally included. The experiments performed in this section using merely SVM fusion are thus previously min-max normalised.

Unimodal systems

First of all, each individual score must be recalculated for the new chimerical multimodal database. Table 4.2 shows the EERs obtained for each prosodic feature using the new data.

Prosodic feature	EER (%)
Log (frames/word)	31.0
Fraction of voiced segments	30.8
Fraction of unvoiced segments	30.8
Log (mean F0)	**21.1**
Log (max F0)	21.2
Log (min F0)	22.4
Log (range F0)	25.9
F0 slope	39.0
Stylised F0 slope	31.5

Table 4.2: EER for each prosodic feature using the multimodal database.

On the other hand, the EERs obtained in each unimodal recognition system are shown in Table 4.3. Note that fusion is only used in the prosodic system, where there are nine prosodic scores to be combined. In this case, fusion is carried out in one single step, and the results for two types of fusion used are presented: matching score fusion with z-score normalization, and support vector machines. On the contrary, no fusion was involved in the unimodal voice spectral and facial recognition systems. It can be seen that the performance of matcher weighting fusion is slightly worse than the performance obtained by applying support vector machines.

Source	Fusion	EER (%)
Prosody	ZS-MW	15.66
	SVM	14.65
Voice spectrum		11.01
Face		2.08

Table 4.3: EER for each unimodal recognition system.

Bimodal systems

When fusing the scores of a bimodal system that contains the prosodic modality, fusion can be carried out in two or one steps, depending on whether the prosodic scores (P) have been previously fused or not, respectively, before fusing them with the spectral scores (S). A diagram of both step strategies is shown in Figure 4.3, which also includes a diagram for the bimodal spectrum-face-based system.

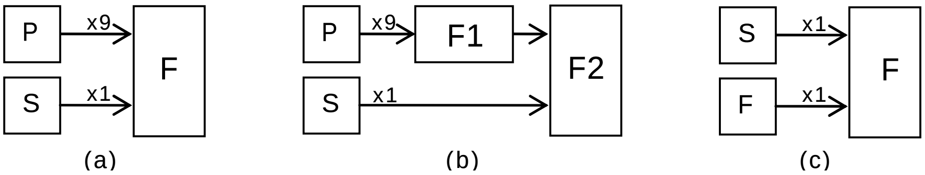


Figure 4.3: Bimodal prosodic-spectral system fused in (a) one step and (b) two steps, and bimodal spectral-facial system (c).

Table 4.4 shows the fusion results for two bimodal systems: a prosody-based system fused with the voice spectral recognition system, and a voice spectrum-based system fused with the facial recognition system, using the same fusion techniques as above (matcher weighting and support vector machines). The prosodic-spectral bimodal system is fused in two different ways. When using only one fusion step, the spectral score and the nine prosodic scores are fused at once. On the other hand, when using two-step, the nine prosodic scores are previously fused, and the resulting final score is fused with the spectral score. In this case, the same fusion technique (MW or SVM) is applied in both steps.

Source	Fusion steps	ZS-MW	SVM
Prosody + voice spectrum	one (a)	7.44	6.84
	two (b)	7.14	6.65
Voice spectrum + face		1.83	0.99

Table 4.4: EER (%) for each bimodal recognition system, using ZS-MW and SVM techniques and one- and two-step fusion strategies in the prosodic-spectral system.

As in the unimodal systems (Table 4.3), SVM fusion outperforms matcher weighting technique. Moreover, in the bimodal prosodic-spectral system the achieved results are better when the two-step strategy is used —i.e. when prosodic scores are previously fused before performing the fusion with the spectral scores—, specially in SVM fusion.

Trimodal system

In a similar manner than bimodal systems, a trimodal system consisting of prosodic, spectral and facial features can be fused in three different ways, according to the number of steps used —one, two or three— in the score fusion.

One-step fusion (Figure 4.4) consists in fusing at once all the scores obtained from the eleven extracted features: nine prosodic scores (P), voice spectral scores (S) and face scores (F). The EERs obtained by using both types of fusion (SVM and MW with z-score normalisation) are shown in Table 4.5.

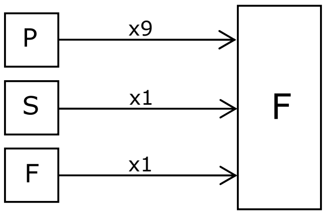


Figure 4.4: One-step fusion strategy for a trimodal system.

Fusion	EER (%)
ZS-MW	1.320
SVM	0.840

Table 4.5: EERs for a trimodal system in one-step fusion using MW and SVM fusions.

The results show, once again, that SVM technique outperforms the conventional MW method with z-score normalisation. Furthermore, by using prosodic features, the results of the bimodal spectral and face recognition system are clearly improved.

Two-step fusion (Figure 4.5) consists in fusing all the scores obtained from each of the eleven parameters in two consecutive steps. In this kind of fusion two different configurations can be considered. In the first configuration (Configuration A), the scores of all the speech features (nine prosodic plus one spectral features) are previously fused and the obtained results are then fused again with the facial scores. In the second configuration (Configuration B), the scores of the nine prosodic features are previously fused and the obtained scores are then fused with voice spectral and facial scores.

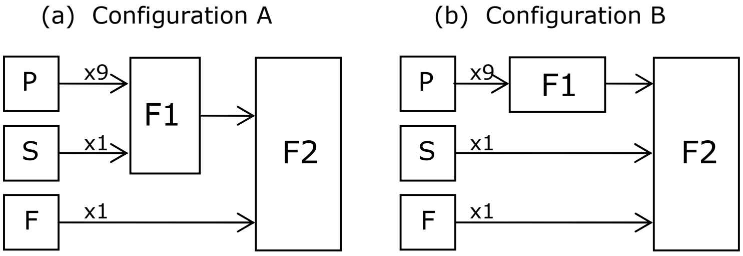


Figure 4.5: Two-step fusion strategies for a trimodal system.

Table 4.6 shows the EERs for both configurations of the proposed two-step fusion. It can be seen that SVM outperforms, once again, the conventional z-score technique. In fact, the best results in both configurations are achieved when SVM is used in both F1 and F2, and the worst results are achieved when using MW in both steps.

F1	F2	Config.A	Config.B
ZS-MW	ZS-MW	2.054	1.493
ZS-MW	SVM	1.880	0.785
SVM	ZS-MW	1.583	1.303
SVM	SVM	0.987	**0.647**

Table 4.6: EERs (%) for a trimodal system in two-step fusion.

Furthermore, applying the two-step Configuration B gives, in all cases, better results than applying Configuration A. The difference is that, when using different fusion techniques in the two fusion steps, the combination SVM[F1] - ZS-MW[F2] performs better in Configuration A, while ZS-MW[F1] - SVM[F2] is better in Configuration B. It seems thus that fusion by SVM is more effective when different modalities are involved.

Three-step fusion In this fusion strategy, scores related to all nine prosodic features are first fused at once. The obtained scores are then fused with voice spectral scores and, finally, the new scores are fused with the facial scores (Figure 4.6).

EERs using ZS-MW and SVM in three-step fusion are shown in Table 4.7. In this case, the same type of fusion is used in all three steps. As in the previous strategies, MW fusion is clearly outperformed by SVM fusion.

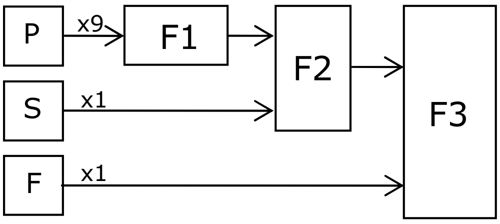


Figure 4.6: Three-step fusion strategy for a trimodal system.

F1, F2, F3	EER (%)
ZS-MW	1.648
SVM	0.868

Table 4.7: EERs for three-step ZS-MW and SVM trimodal fusions.

4.4.2 HE-SVM multimodal fusion

First of all, a 1000-interval histogram equalisation is applied to unimodal (prosodic) and bimodal systems, in order to see how this normalisation technique performs over SVM fusion. The scores are always equalised to the best score distribution involved in the fusion process. Table 4.8 shows the results obtained when fusing the prosodic scores and the bimodal systems prosody-spectrum and spectrum-face with HE-SVM technique. When using a two-step fusion in the prosody-spectrum system, HE-SVM is used in both steps. Simple SVM is also shown for comparison.

Except for the bimodal spectrum-face system, where similar results are achieved by using SVM and HE-SVM, the performance is clearly improved by applying HE to SVM fusion. However, the best relative improvement can be seen in the two-step prosody-spectrum fusion, where HE-SVM is applied twice.

Source	Fusion steps	SVM	HE-SVM
Prosody		14.65	13.39
Prosody + voice spectrum	1	6.84	6.25
	2	6.65	5.55
Voice spectrum + face		0.99	1.02

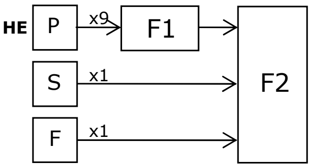
Table 4.8: EER (%) for the prosodic and each bimodal recognition system using SVM and HE-SVM techniques.

In order to analyse how the trimodal fusion process is influenced by a previous histogram equalisation of the scores, HE is applied to the different fusion strategies presented above. First, histogram equalisation is applied to trimodal fusion in one-step (as shown in Figure 4.4). As well as in the bimodal fusion, all the scores are equalised to the best score distribution —the facial distribution, in this case. The obtained results can be seen in Table 4.9.

Second, HE is applied to the fusion strategy achieving the best results in the previous section, i.e. configuration B in two-step fusion (Hernando et al., 2006). In this case, HE is applied before the first fusion step (F1) and all the score distributions are equalised to the best score distribution. Since only prosodic scores are involved in F1, all the scores are equalised to the *mean F0* distribution.

Fusion	**EER (%)**
SVM	0.840
HE-SVM	0.680

Table 4.9: EERs for a trimodal system in one-step fusion using SVM and HE-SVM techniques.



F1	F2	no HE	HE
ZS-MW	ZS-MW	1.493	0.987
SVM	ZS-MW	1.303	0.886
ZS-MW	SVM	0.785	0.774
SVM	SVM	0.647	0.630

Table 4.10: EERs (%) with equalised and non-equalised scores in the best fusion strategy.

The results obtained with equalised and non-equalised scores are shown in Table 4.10, where it can be clearly seen that a previous histogram equalisation as a normalisation technique improves, in all cases, the results obtained with non-equalised scores. Moreover, the relative improvement is more considerable where MW is used in the second fusion (F2), i.e. where the non-equalised fusion had the worse results.

In Table 4.10, HE was only applied before the first fusion step (F1). However, in this two-step trimodal configuration, HE can be used in three different ways, depending on whether equalisation is applied only before F1, only before F2, or both, as listed below (Farrús et al., 2007):

(1) HE before the first fusion step (equalisation of the prosodic scores);

(2) HE before the second fusion step (equalisation of all three modalities); and

(3) HE before both fusion steps F1 and F2.

These three methods are now applied over the last case in Table 4.11, which uses SVM fusion in both F1 and F2. The results obtained are shown in Table 4.11, where they can be compared with the non-equalised SVM fusion.

F1	**F2**	**EER(%)**
SVM	SVM	0.647
HE-SVM	SVM	0.630
SVM	HE-SVM	0.649
HE-SVM	HE-SVM	**0.613**

Table 4.11: EERs (%) obtained applying HE before trimodal two-step SVM fusion.

As it can be seen in the table, the best result is achieved when histogram equalisation is used before F1 and F2. By equalising only the prosodic scores the performance of the system is also improved. On the other hand, equalisation before the second fusion does not improve the performance of the overall system. This contrasts with the results of Configuration B in Table 4.6, where SVM fusion was more effective when used in F2 than when used in F1, and it suggests that equalisation is more effective when it is performed on the scores with the highest values of EER (the prosodic scores in this case).

Finally, several equalisation combinations can be performed in a trimodal three-step fusion depending on the place where HE is performed. In this section, three different equalisation combinations have been tested: HE before the first fusion step (F1), HE before the last fusion step (F3), and HE before each fusion step (F1, F2 and F3). Table 4.12 contains the EER obtained in each combination. The non-equalised three-step SVM fusion is also shown for comparison.

F1	**F2**	**F3**	**EER(%)**
SVM	SVM	SVM	0.868
HE-SVM	SVM	SVM	0.643
SVM	SVM	HE-SVM	0.649
HE-SVM	HE-SVM	HE-SVM	**0.624**

Table 4.12: EERs (%) obtained applying HE before three-step SVM fusion.

The use of HE-SVM fusion in at least one of the fusion steps contributes —to a large extent— to improve the non-equalised SVM. However, the best performance is clearly obtained when HE is applied to each fusion step. Equalisation of the prosodic scores (HE before F1) and equalisation of the three modalities (HE before F3) give comparable results, but still the former gives a slightly better result than the latter. This fits in with the results obtained in the trimodal two-step fusion (Table 4.11), where the best result was obtained by equalising all the fusion steps, and equalising only the first step was better than equalising only the last step.

4.4.3 MW, SVM and HE-SVM using a reduced database

The experiments performed up to now made use of the multimodal database described in section 4.3. Nevertheless, the dimension of the database was considerably large, so that the computational cost of the experiments —specially when using SVM— was very high. In order to test the performance with other fusion and normalisation techniques, a reduced multimodal database was constructed.

Experiments using other fusion techniques will not be presented in this thesis. However, this new reduced database will be used to correct the error committed in the prosodic parameters. As it was explained in section 4.1, the second and third prosodic parameters —fraction of voiced and unvoiced segments, respectively— are completely correlated, so that the use of one of the parameters is totally redundant to the other one, so that their EERs are the same.

In this section, the *fraction of unvoiced segments* parameter will be removed, and the most significative multimodal fusion experiments will be performed again using only eight prosodic parameters —instead of nine— in the reduced database. In order to know the influence of the reduction of the database dimension on the final results, experiments using nine parameters in the new database will be also presented.

The multimodal reduced database has been built in the same way as the chimerical database described in section 4.3, i.e. combining the Switchboard-I speech database and the video part of the XM2VTS database. The new multimodal database consists of the same amount of users, but the number of scores corresponding to the development set was reduced from 46500 to 930 (336 client trials and 594 impostor trials), while the number of scores of the test set was maintained in 46500 (16800 client trials and 29700 impostor trials). Since the training set has less than 1000 scores, a 100-interval histogram equalisation has been applied.

Unimodal systems

First of all, each individual score must be recalculated using the reduced mutimodal database. Table 4.13 shows the EERs obtained for each prosodic feature using the new data.

Feature	EER (%)
Log (frames/word)	32.1
Fraction of voiced segments	29.8
Log (mean F0)	**19.0**
Log (max F0)	20.9
Log (min F0)	22.2
Log (range F0)	27.8
F0 slope	38.0
Stylised F0 slope	28.6

Table 4.13: EER (%) for each unimodal recognition system.

On the other hand, the EERs obtained in each unimodal recognition system are shown in Table 4.14. Fusion is only used in the prosodic system and it is performed in one single step using ZS-MW, SVM and HE-SVM techniques. In order to allow the comparison with the large database, the prosodic system uses eight (Prosody 8) and nine (Prosody 9) parameters.

Source	Fusion	EER (%)
Prosody 9	ZS-MW	15.66
	SVM	14.88
	HE-SVM	13.69
Prosody 8	ZS-MW	14.88
	SVM	14.88
	HE-SVM	14.29
Voice spectrum		9.52
Face		2.50

Table 4.14: EER for each unimodal recognition system.

Trimodal systems

The reduced database is also used to test the trimodal prosodic-spectral-facial system by means of the most significant step strategies. The results, summarised in Table 4.15, allow to compare the performance of the system when using nine and eight prosodic scores together with the spectral and facial scores. Only configuration B (Figure 4.5) is used in the two-step strategy, and for each strategy, all three ZS-MW, SVM and HE-SVM techniques are applied.

Steps	Pros. features	ZS-MW	SVM	HE-SVM
1	9	**1.417**	1.232	0.989
	8	**1.428**	1.267	0.893
2 (config. B)	9	1.673	**1.071**	**0.660**
	8	1.690	1.226	0.717
3	9	1.994	1.185	0.738
	8	2.006	**0.991**	**0.673**

Table 4.15: EERs (%) for a trimodal system using ZS-MW, SVM and HE-SVM techniques in one, two and three fusion steps over the reduced database.

The use of nine prosodic characteristics instead of eight in ZS-MW fusion could be interpreted as giving an additional weight to one of the eight prosodic features (namely the *fraction of voiced segments*). The EER of this feature in the development set equaled 30.88%, a value that was greater than the mean of the prosodic features, which equaled 27.86%. Therefore, since the weight assigned to this feature was reduced when using eight prosodic features, and since the performance of such feature was worse than the mean, it would be expected *a priori* to get better results when using eight prosodic features. This was the case in the ZS-MW unimodal prosodic system (Table 4.14), but not in the trimodal system (Table 4.15), where the performance achieved with nine characteristics is always slightly better when using ZS-MW fusion. This could be explained by the fact that, when reducing the weight of *fraction of voiced segments*, the weights assigned to the other features are automatically increased. Most of these features perform better than the *fraction of voiced segments*; however, some of them have higher EERs, and more weight is also given to them, which may cause worse final results. Nevertheless, the differences between the use of nine and eight prosodic characteristics do not seem to be significant, which leads to believe that the redistribution of the weights may be rather balanced.

In the case of SVM and HE-SVM, no significant differences seem to appear depending on the number of prosodic features used, neither. Whereas using 8-prosodic features database gives better results in the three-step and in the HE-SVM one-step trimodal strategies, the fusion is outperformed over the 9-prosodic features database in the other cases (except for the unimodal SVM fusion, where the same result is observed). Thus, removing the correlated feature does not modify SVM and HE-SVM in a consistent direction: the fusion performance is sometimes improved and sometimes decreased.

This comparison between the use of 8 and 9 prosodic features in the reduced database leads to believe that, probably, if the correlated feature would have been removed in the extended database, the new results would not differ to a large extent from those already obtained with nine prosodic features. In order to check statistically the significance between both 8- and 9-prosodic feature databases, a statistical hypothesis t-test has been conducted. The t-test gives the probability that the difference between two values is caused by chance. Such test is basically valid for testing the difference between two means and not between proportions; however, the t-test in proportions has been extensively studied and found to be robust, except if one of the proportions is very close to zero, one or minus one.

The t-test has been performed by choosing a 95% confidence interval (i.e., a significance level of 0.05) and computed over all the step strategies of ZS-MW, SVM and HE-SVM fusions. In most cases —except in two- and three-step SVM fusion— the doublesided p-value was greater than 0.05, which means that the differences between most of the proportions in 8- and 9-prosodic databases is not significant. This corroborates statistically the *a priori* hypothesis: using eight prosodic features in the extended database would not give significant differences in comparison with the results obtained using nine prosodic features, at least in ZS-MW and HE-SVM fusions.

Larger differences in the results are noticed when applying different step strategies. Fusing all three modalities in one step, for instance, give the best results in ZS-MW but the worst ones in SVM and HE-SVM techniques. A possible explanation could be the following: the two- and three-step strategies were initially proposed because a high degree of correlation within the speech features —specially within most of prosodic features— was assumed. Thus, these step fusion strategies tried to group those features that were correlated before fusing them. Whereas SVM fusion *learns* about such correlation and takes advantage of this feature grouping, MW does not; the feature correlation is assumed but it is not considered in the MW technique. Therefore, unlike SVM fusion, it is better not to group (i.e. to assume correlation) within features in a MW fusion process.

Both SVM and HE-SVM perform much better in two- and three-step fusions. Specifically, the EERs in two-step fusion are lower when using all the features, while in three-step fusion the EERs are lower when using only eight prosodic features together with the spectral and facial ones. However, the differences between two and three steps are not considerable enough to reach a satisfactory conclusion.

4.5 Conclusions

The performance of a bimodal system based on facial and spectral information is clearly improved by adding prosodic information to the system. However, the results vary depending on the fusion technique ustilised. In these experiments, the use of SVM fusion outperforms the results obtained with MW technique.

Moreover, the way how the scores are fused is relevant for the performance of the system. When using ZS-MW in the trimodal system, the best results —both in the large and the reduced databases— are obtained in the one-step fusion strategy, probably due to the fact that the correlation assumed when grouping the features in the other step strategies is not taken into account in the MW technique. On the contrary, SVM technique performs normally better in the Configuration B of the two-step fusion, except when 8 prosodic parameters are used in the reduced database, where the best result is achieved in the three-step fusion. However, in the experiments performed over the extensive database it can also be observed that a previous fusion of the voice information —spectral and prosodic scores— does not contribute, in any case, to the improvement of the system.

In addition, results are improved by applying histogram equalisation as a normalisation technique. In these cases, the performance ranking for the different strategies is normally maintained with respect to the use of SVM without HE. In HE-SVM, the score distribution is always equalised to the best score distribution involved in the fusion. *A priori*, this was done in this way for a common sense motivation; however, although it has not been demonstrated in this thesis, some experiments performed over the same database have shown that this is the way to obtain better fusion results.

The most significant results obtained in this chapter are summarised in Table 4.16. Obviously, the best results are obtained over the large database; however, the final conclusions are similar in both databases. Moreover, no significant differences appear between using 8 or 9 prosodic parameters in the reduced database, which leads to the hypothesis that using 8 prosodic parameters in the large database would neither contribute with significant differences.

Margin errors have been computed for the best results obtained in the extensive database, leading to 1.320 ± 0.100, 0.647 ± 0.073, and 0.613 ± 0.067 for 1-step ZS-MW, 2-step SVM and 2-step HE-SVM, respectively. Although certain overlapping can be observed between SVM and HE-SVM, there is a clear tendency to obtain, in all cases, lower EERs in HE-SVM fusion. Other confidence intervals have been calculated by splitting the test set in 30 small sets and computing the average and standard deviation of such 30 EERs values, using also a 95% Wald confidence interval for an average value. Using this sampling design, the confidence intervals obtained were similar —or even wider— to the ones shown above.

Database	Steps	ZS-MW	SVM	HE-SVM
Extensive	1	**1.320**	0.840	0.680
	2	1.493	**0.647**	**0.613**
	3	1.648	0.868	0.624
9-Reduced	1	**1.417**	1.232	0.989
	2	1.673	**1.071**	**0.660**
	3	1.994	1.185	0.738
8-Reduced	1	**1.428**	1.267	0.893
	2	1.690	1.226	0.717
	3	2.006	**0.991**	**0.673**

Table 4.16: Summarised results for both extensive and reduced databases, using one-, two- and three-step strategies and ZS-MW, SVM and HE-SVM fusion techniques.

Chapter 5

Robustness Analysis to Imitated and Converted Voices

Voice imitation and other types of disguise are potential threats to security systems that use automatic speaker recognition; therefore, several studies have been performed in order to test the vulnerability of speaker recognition systems against imitation by human or synthetic voices. The main objective of this chapter is to analyse human voice imitations and synthetic converted voices in order to know how vulnerable auditory and automatic speaker recognition systems are to this kind of disguise.

First, some experiments are performed in section 5.1 in order to test the influence of foreign accents and dialects —as a sort of imitation— in auditory speaker recognition. In section 5.2, the voices of two professional imitators trying to impersonate several well-known politicians are used to analyse the behaviour of some selected prosodic and acoustic features in the imitated voices. Finally, automatic converted voices are used in section 5.3 in order to test the robustness of a speaker identification system against this kind of synthetic speech. Conclusions of the chapter are presented in section 5.4.

5.1 Dialect and accent imitation

As it was said above, voice disguise is a possible threat to the performance of a speaker recognition system and to the accuracy of earwitness descriptions. One common disguise is, for instance, the modification of the own dialect or accent.

Stage dialects are used in the theatres and film. This sort of dialect imitation could however be used in a criminal setting, for example, and for this reason it is relevant to investigate how convincing dialects imitated for theatrical use are. In these section, this kind of disguise is explored using recordings from a well-known English-speaking actor with considerable experience of dialect and accent imitation.

The first approach of the investigation presented is the human perception element. In this work, it is considered whether theatrical accent modification affects human perception in the process of voice discrimination. In order to see how successful dialect imitations are and how the process of speaker discrimination is influenced by accent disguise, two sets of human perception tests are constructed and presented in sections 5.1.1 and 5.1.2, respectively. The first set is focused on American and British English dialects, and the second set on American and London English- and Spanish-accented English. Each set consists of three parts: a same-different speaker test, a same-different accent test, and a select the accent from a closed-set of options test.

A second approach investigates whether speaker recognition systems are vulnerable to dialect and accent disguise. Since dialectal imitations and disguises are based mainly on what is usually called high-level characteristics such as intonation and lexical terms, it is not expected *a priori* to find a spectral automatic recognition system capable of classifying one speaker's talk according to the accent spoken. However, frequency characteristics can be affected by modification in intonation and acting. In section 5.1.3, the same speech segments used in the perception test were used in an automatic speaker recognition experiment in order to compare the results and to check the robustness of the system in front of the voice changes.

Finally, the results and conclusions of these experiments (presented in Farrús et al. (2006a)) are shown in sections 5.1.4 and 5.1.5, respectively.

5.1.1 Perception experiment on English dialects

Participants

Thirty-six males and twelve females aged between 14 and 54 years who were native speakers of English or judged themselves to be advanced learners took part in the experiment. None of the participants reported a hearing problem. The participants were recruited by the experimental leaders and requests to friends to spread information about the experiment.

Material

From the following movies and extra material available on DVD, five three-four-second segments were selected; for Speaker A: *Chocolat* (Irish), *Blow* (American), *Finding Neverland* (Scottish), *From Hell* (British English), *Secret Window* (American), and Speaker A's own voice (extra material), and for Speaker B: another male American actor speaking General American.

Speaker	Dialect
A	Irish
	American
	Scottish
	British English
	American
	American (own voice)
B	American (own voice)

Table 5.1: Speakers and their corresponding dialects involved in Experiment 1.

Procedure

The participants undertook three tests in a web-browser environment. Before Test 1, demographic, hearing, and language competence data was collected. In Test 1 each participant was presented with 15 randomly constructed pairs of stimuli selected from the 25 available stimuli. They were given the instructions ”You are going to hear a set of files. Each file contains *two* passages. Your task is to decide if they are passages spoken by the same speaker. Indicate your decision by selecting the circle *Yes* or *No*”.

In Test 2 each participant was presented with 15 randomly constructed pairs of stimuli selected from the 25 available stimuli. They were given the instructions ”You are going to hear a set of files. Each file contains *two* passages. Your task is to decide if the passages are spoken by speakers of the same dialect/accent/regional variety or not. Indicate your decision by selecting the circle *Yes* or *No*”.

In Test 3 each participant was presented with 15 randomly selected stimuli from the 25 available stimuli. They were given the instructions, ”After having listened to each file you are asked to choose from the drop-down list [American, English, Scottish, Irish, Welsh, Australian, New Zealand, South African, Spanish] which accent the speaker has”.

5.1.2 Perception experiment on accents and dialects

Participants

Ten males and seven females aged between 21 and 60 years who were native speakers of Spanish or judged themselves to be advanced learners took part in the experiment. None of the participants reported a hearing problem. The participants were recruited by the experimental leaders and requests to friends to spread information about the experiment.

Material

The Scottish and the Irish accents of Experiment 1 were exchanged for Speaker A’s Spanish-accented English (*Before Night Falls*) and Speaker C, a Spanish male actor speaking English with a strong Spanish Accent.

Speaker	Dialect/accent
A	American
	Spanish
	British English
	American
	American (own voice)
B	American (own voice)
C	Spanish

Table 5.2: Speakers and their corresponding dialects and accents involved in Experiment 2.

Procedure

The procedure was identical to Experiment 1.

5.1.3 Automatic speaker recognition experiment

The second approach investigates whether speaker recognition systems are vulnerable to dialect and accent disguise. Since dialectal imitations and disguises are based mainly on phonetic, intonation and lexical characteristics, it is not expected *a priori* to find a spectral automatic recognition system capable of classifying one speaker's talk according to the accent spoken. However, frequency characteristics can be affected by modification in intonation and acting. In this section, the same speech segments used in the previous experiments are used in an automatic speaker recognition experiment order to compare the results and to check the robustness of the system in front of the voice changes.

Two testing evaluations were designed. For both conditions, the speech files were parameterised using 20 filterbanks to form 20 MFCC for a frame size of 24 ms and a shift of 8 ms. Delta and acceleration coefficients were included. Speaker GMMs were used, formed from 32 Gaussian mixture components and trained with four of the five speech segments. Five tests were realised for each voice/dialect by alternating the training and test files: each time a different subset of four files was used for training with the remaining fifth being used for testing.

In Evaluation Condition 1, Speaker A's natural voice, his imitated dialects and the voices of the Speakers B and C were used to train different speaker models. The same set of speakers and imitated dialects were used in the test phase. In Evaluation Condition 2, the set of speaker models was reduced to three: Speaker A's natural voice, Speaker B and Speaker C. Speaker A's imitated dialects were used for testing.

5.1.4 Results

The results of the human perception Tests 1 and 2 for Experiments 1 and 2 are shown in Table 5.3. The results of Test 3 for Experiments 1 and 2 are shown in Table 5.4. A_OAm indicates speaker A's own American accent, A_IEng indicates speaker A's Imitation of an English accent, A_ISc indicates speaker A's Imitation of a Scottish accent, A_IIr indicates his imitation of an Irish accent and A_ISpa his imitation of Spanish accent-English. B and C indicate speakers B and C, respectively.

The results of the Automatic Speaker Recognition experiments are shown in Table 5.5 (Condition 1) and Table 5.6 (Condition 2).

5.1.5 Discussion and conclusion

The results of Experiments 1 and 2 show that the actor (Speaker A) was successful in his dialect imitations and that these accents would result in a witness describing his accent as one other than his native. The results also suggest an interaction between accent and knowledge of the accent —or accents— in focus. This is seen in the difference between the Experiment 1 and 2 listeners' responses to the imitated English accent (Table 5.3), and the inability of the

	A_OAm		A_IEng		A_ISc		A_IIr		B	
	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
A_OAm	73	81								
A_IEng	13	*19*	84	90						
A_ISc	11	*6*	30	*22*	70	75				
A_IIr	17	*28*	53	*50*	26	*25*	73	75		
B	*32*	62	*6*	*13*	*3*	*6*	*5*	*16*	62	87

(a)

	A_OAm		A_IEng		A_ISpa		B		C	
	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
A_OAm	71	80								
A_IEng	35	*33*	76	67						
A_ISpa	0	*7*	12	*7*	82	87				
B	*35*	67	0	*53*	*6*	*13*	76	87		
C	*6*	*7*	*0*	*13*	*18*	40	*0*	*13*	59	67

(b)

Table 5.3: Percentage of Yes-responses in Experiments 1 (a) and 2 (b) for Test 1 (T1) (Speaker) and Test 2 (T2) (Dialect). Italics indicate when the correct answer was *No*.

	Response								
Stimulus	**Am**	**Eng**	**Sc**	**Ir**	**Spa**	**SA**	**Aus**	**NZ**	**Welsh**
A_OAm	*79*	2	0	2	1	1	9	4	1
A_IEng	3	*59*	5	3	1	3	10	14	1
A_ISc	0	7	*44*	37	1	3	2	2	3
A_IIr	18	19	10	*15*	0	5	10	6	17
B	*85*	8	2	0	0	2	2	0	0

(a)

	Response								
Stimulus	**Am**	**Eng**	**Sc**	**Ir**	**Spa**	**SA**	**Aus**	**NZ**	**Welsh**
A_OAm	*83*	3	0	0	0	0	10	5	0
A_IEng	23	*49*	9	6	0	3	6	3	3
A_ISpa	5	8	8	5	*49*	18	3	3	3
B	*83*	11	1	1	0	0	1	0	3
C	0	5	0	3	*63*	28	3	0	0

(b)

Table 5.4: Percentage dialect selection (test 3) by Experiment 1 (a) and 2 (b) listeners. Italics indicates correct dialect selection where correct is defined as the dialect Imitated for the A_I voices and the speakers actual dialect for the A_Oam, B and C voices.

	Assigned voice						
Test data	**A_OAm**	**A_IEng**	**A_ISc**	**A_IIr**	**A_ISpa**	**B**	**C**
A_OAm	4	0	0	0	1	0	0
A_IEng	0	5	0	0	0	0	0
A_ISc	0	0	1	2	0	1	1
A_IIr	0	0	0	5	0	0	0
A_ISpa	1	0	1	0	3	0	0
B	0	0	0	0	0	5	0
C	0	0	0	0	0	0	5

Table 5.5: Voice assignment from the automatic speaker recognition condition 1 test, where each of Speaker A's dialects formed a separate speaker model.

	Assigned voice		
Test data	**A_OAm**	**B**	**C**
A_IEng	0	1	4
A_ISc	0	1	4
A_IIr	0	0	5
A_ISpa	3	0	2

Table 5.6: Voice assignment from the automatic speaker recognition condition 2 test, where none of Speaker A's dialects formed a training speaker model.

Experiment 1 listeners to deal with the imitated Irish accent (Table 5.3a and Table 5.4a). A similar —though much weaker— trend is seen in for the imitated Scottish voice. The difference could be due to Irish imitation being of poorer quality that the Scottish imitation or that more of the listeners were familiar with the Scottish accent. This can only be resolved by running the experiment groups of Scottish and Irish listeners.

Yet, in spite of the variation in the familiarity with the different accents, when same dialect and speaker segments were presented (T1), the yes response rate was markedly stable. For Experiment 1, the strongest yes responses were for the imitated-English accent and for Experiment 2, the Spanish-accented English accent. However, as can be seen in Table 5.4, the listeners were less able to place these imitated accents; here the real American accents are the best placed. Of note is that Table 5.4b reveals that the well-known actor (Speaker A) has picked up on an aspect of the Spanish-accent English accent that led the respondents to select South Africa (SA) paralleling the responses to the native Spanish speaker.

The dialect task (T2) revealed that, with the exception of the responses to the imitated English accent by the Experiment 2 listeners, the recognition of the same dialect outscores recognition of the same speaker. Hence, it appears that accent-specific features dominate speaker-specific features (Eriksson et al., 2007).

The automatic speaker recognition experiment that used individual models for each of the imitated dialects (Evaluation Condition 1) resulted in little confusion between models. Further, only the imitated Scottish accent resulted in confusion between speakers A, B and C. Speakers B and C were perfectly recognised. However, under Evaluation Condition 2, where the three speaker models used were Speaker A's American Accent, Speaker B and Speaker C, and the imitated accents used as test data, only 3 of the 20 samples were correctly identified. These were imitated Spanish-accent English segments. This condition suggests that the Speaker A's dialect imitations (in the native language) include changes that affect automatic speaker recognition systems that Spanish accented-English imitation does not. In his Spanish accented English (where Spanish is a foreign language) he may have concentrated on other aspects of the voice, such as the feature that result in the selection of South Africa as the dialect (Table 5.4), that perhaps could never result in this imitated voice becoming distinct from his natural voice.

Together the experiments presented in this paper show that dialect imitation can confuse both the human listener and speaker recognition systems, yet in different ways, and that high-quality dialect disguise is a topic that warrants forensic linguistic consideration.

5.2 Vulnerability of prosodic and acoustic features to voice imitation

Voice imitation is one of the potential threats to security systems that use automatic speaker recognition. Since prosodic features have been considered for state-of-the-art recognition systems in recent years, the question arises as to how vulnerable these features are to voice mimicking. In this section, two experiments are conducted for twelve individual features in order to determine how a speaker identification system would perform against professionally imitated features. Such features include nine prosodic parameters and three voice quality (acoustic) parameters based on the prosodic and *JitShim* systems presented in chapter 3.

The current study explores the ability of professional mimickers to approximate the source parameters and prosody of their target voices. The study comprises a set of experiments, in which professional voice imitators mimic the voice characteristics of well-known public figures. In each experiment, typical source-related parameters are measured and compared between the target speaker's voice (*target*), the imitator's natural voice (*i-natural*) and the imitator's modified voice (*i-modified*).

Twelve source- and prosody-related parameters include the length of words and word segments, means, extrema and ranges of the fundamental frequency, and jitter and shimmer. For each of those parameters, a baseline speaker identification experiment is conducted to establish the error rate in per cent of a speaker identification system that would try to distinguish between the target speaker and the imitator's natural voice on the basis of the single source parameter. Then, a second experiment is conducted —again for each individual source parameter— to establish the error rate in per cent of a speaker identification system that would try to distinguish between the target speaker and the imitator's modified voice, again on the basis of the single source parameter. It is the comparisons between the two experiments that reveal, for each of the twelve parameters, how much the professional imitator is able to shift the parameter away from his own voice towards the target speaker's voice. In turn, these comparisons establish the vulnerability of the twelve source parameters against intentional voice mimicking by professionally trained impersonators.

5.2.1 Material

Two male professional imitators, who will be referred to with their initials (cc and qn) took part in these experiments. They have been working as professional imitators on radio and TV for more than five years. They both are Catalan native speakers and have a Central Catalan dialect.

Five male well-known politicians, who will be referred to with their initials (JB, JR, JS, PM and XT) were used as target speakers. They were between 45 and 64 years old when the recordings were made. JS, PM and XT are Catalan native speakers from the same dialectal region as the professional impersonators, while the remaining two (JB and JR) are Spanish native speakers with a Castilian Spanish dialect. Table 5.7 summarises the main characteristics of the impersonators and the target speakers.

	Speaker	Age	Language	Dialect
Impersonators	qn	44	Catalan	Barcelona
	cc	32	Catalan	Barcelona
Target politicians	PM	66	Catalan	Barcelona
	XT	60	Catalan	Barcelona
	JR	46	Spanish	Valladolid
	JS	57	Catalan	Barcelona
	JB	56	Spanish	Albacete

Table 5.7: Characteristics of the impersonators and the target politicians.

The recording of the target speakers were taken from public radio interviews, made in local radio station's studios. For each target voice, 20 sentences of about 10-20 seconds length were extracted. The imitations and the natural voices of the impersonators were recorded in their own radio station's studio or in an audio studio at the Department of Signal Theory and Communications at Technical University of Catalonia (Figure 5.1). The advantage of having recordings from radio interviews and the corresponding imitations made in closed studios without video cameras is that, when the impersonator is not seen by the audience, it is more important to focus on voice similarity, since the listener has no clues other than the voice and speech to identify the target speaker (Zetterholm, 2003).

Figure 5.1: Recording session at Technical University of Catalonia.

The impersonators were asked to record both imitated and natural voices with the same text as the recordings of the target speakers. Since a read-text recording may result in a lack of spontaneity, the impersonators had been reading the texts before the recordings in order to copy the target voices as naturally as possible. The impersonator qn imitated the politicians JR, PM and XT, and cc imitated JB and JS. Table 5.8 shows imitators and target speakers together with the mean fundamental frequency of each speaker. The standard deviation is also shown as a margin error. Both impersonators recorded all the extracted sentences of each target with their natural (*i-natura*l) and modified (*i-modified*) voices. All the transcriptions were manually word-labeled and aligned.

Imitator	F0 (Hz)	Target	F0 (Hz)
cc	121 ± 37	JB	110 ± 44
		JS	85 ± 54
qn	110 ± 23	JB	81 ± 22
		JR	95 ± 67
		XT	87 ± 27

Table 5.8: Mean F0 of impersonators and target voices.

5.2.2 Experimental setup

Both impersonator's voices (*i-natural* and *i-modified* voices) were recorded at the same time and in the same recording conditions, while target voices were extracted from previous radio recordings. Due to this mismatch and the small number of speakers used in the experiments, it was not reliable to perform the recognition task with a conventional cepstral-based GMM method. Therefore, only source- and prosody- related parameters were considered, since they seem to be more robust to mismatched recordings.

For each *i-natural*, *i-modified* and target voice, a vector consisting of the following twelve source- and prosody-related features was extracted to perform the identification experiments:

- Log (number of frames per word)
- Length of word-internal voiced segments
- Length of word-internal unvoiced segments
- Log (mean F0)
- Log (max F0)
- Log (min F0)
- Log (range F0)
- F0 slope
- Mean F0 absolute slope
- Jitter (absolute)
- Shimmer (absolute)
- Shimmer (apq3)

These features are based on the prosodic and *JitShim* systems described in chapter 3. Since the compared voices had the same text, the *fraction of voiced segments* feature used chapter 3 was replaced by features based on the length of voiced and unvoiced segments, which are presumably more text-dependent, and the mean slope of the stylised F0 contour was replaced by the mean F0 absolute slope. Although jitter and shimmer are not normally considered prosodic parameters, all the features below will be referred to, for simplicity, as prosodic features.

The parameters were extracted using the Praat software for acoustic analysis (Boersma and Weenink, 1992), performing an acoustic periodicity detection based on a cross-correlation method, with a window length of 40/3 ms and a shift of 10/3 ms. The mean over all words was computed for each individual feature.

For every set of 20 different sentences, one speaker model was trained for the *i-natural* voice and one for the target voice. Either five or ten sentences were used for training the models. The remaining sentences, together with the corresponding *i-modified* sentences, were used for testing. The system was tested using the k-nearest neighbour classifier (with $k = 1$ and k=3), comparing the Euclidean distances of the test feature vector to the k closest vectors of each set of the trained speaker models.

For each of the twelve parameters, a baseline speaker identification experiment was conducted to establish the error rate of a speaker identification system, which tried to identify the target and *i-natural* voices from the closed set of two speaker models: the mimicker using his natural voice and the corresponding target speaker, both trained using the same set of sentences. Again for each individual parameter, a second experiment was conducted to establish the error rate of an identification system that tried to identify the target and *i-modified* voices from the same closed set of two speaker models: the impersonator speaking with his natural voice and his corresponding target speaker. So, in each identification experiment, a total number of 150 tests were performed when the models were trained with 5 sentences (5 targets x 2 speakers x 15 sentences) and 100 tests were performed when the models were trained with ten sentences (5 targets x 2 speakers x 10 sentences).

Finally, the fusion of all the individual features was performed in each experiment at the score level (see section 2.4.2). The scores were normalised with the well-known z-score normalisation, which transforms the scores into a distribution with zero mean and unitary variance, and fused with the matcher weighting method, where each individual score is weighted by a factor proportional to the recognition rate (Indovina et al., 2003).

5.2.3 Identification results

The identification error rates (IERs) obtained for both baseline and modified systems are presented in per cent in Table 5.9. The baseline system is tested with the *i-natural* and target voices, while the modified system utilises the *i-modified* and target voices for testing. In the modified system, *identification error* means that the *i-modified* voice was identified as the target speaker's voice instead of the imitator's own voice.

The error rates are given for the whole prosodic systems, that is, after fusing all the twelve features involved in the experiments. The table shows the results obtained by using five and ten sentences to train the speaker models. In both cases, the error rates when using $k = 1$ and $k = 3$ in the k-nearest neighbour classification are compared.

Training	**1st NN**		**3rd NN**	
	baseline	modified	baseline	modified
Five sentences	10.3	19.3	8.7	18.3
Ten sentences	5.0	22.0	11.0	**18.0**

Table 5.9: IER (%) obtained for each prosodic system after fusing all the features.

The results clearly show that, after fusing all the features, the identification error is always increased when using the modified system instead of the baseline system. The largest difference can be seen when using the 1st nearest neighbour as a classifier and 10 sentences for training.

The identification error rates for each isolated feature are plotted in Figure 5.2, where the green (light) line corresponds to the IERs of the baseline system and the red (dark) one to the IERs of the modified system. In all cases analysed in Table 5.9, the results for every individual feature were similar; therefore, only one case (the 1st nearest neighbour and 10 sentences for training) is represented in the figure.

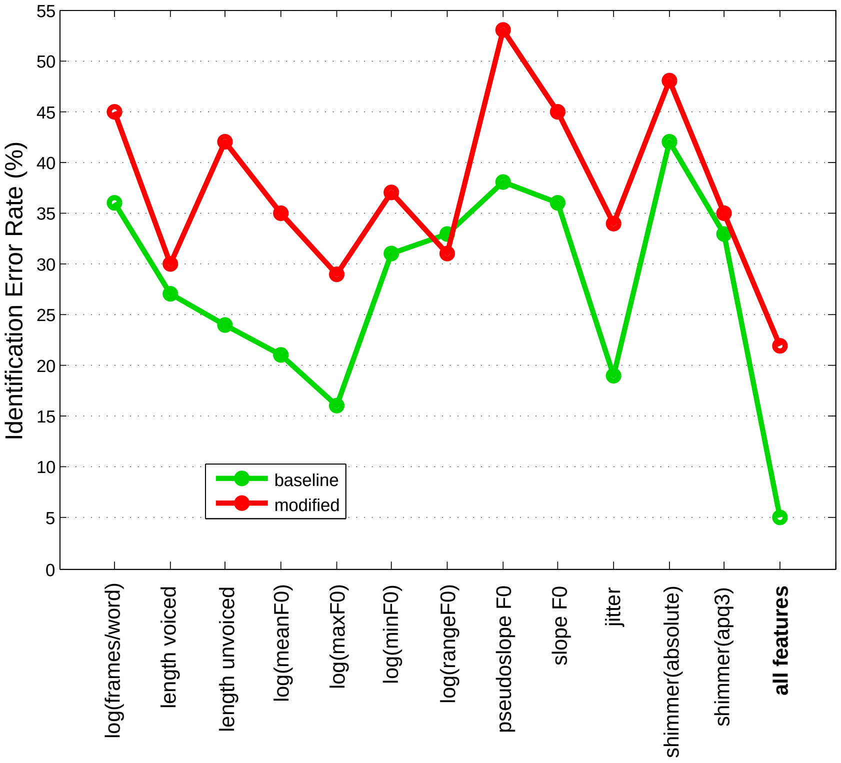


Figure 5.2: IER (%) for each prosodic feature (and fusion) using 1st NN and 10 sentences for training.

As can be seen in the figure, the error rates increase in all the individual parameters except in one: the range of the fundamental frequency (i.e. the difference between the maximum and minimum values of F0), which remains steady —or even decreases, in this case— in the modified system.

5.2.4 Conclusions

A set of experiments was conducted, in which twelve prosodic and source-related features were used for speaker identification, and where a professional impersonator attempts to mimic a target voice. For each individual feature, a baseline experiment established models for the target speaker and the natural voice of the impersonator, using a set of training data. A separate set of test data from the target and the impersonator's natural voice was then used to determine the identification error rate for the two speakers *without* attempted impersonation. For each of the twelve features, a second experiment was then conducted, which used the target speaker's test data and the impersonator's modified voice data to determine the identification error rate for the two speakers *with* attempted impersonation. For eleven of the twelve features, the identification error rate increased, in some cases greatly, but for the F0 range the identification error rate remained almost unchanged —in fact even dropped slightly from 33% to 31%. Fusing the twelve features at the score level resulted in an increase from an identification error rate of 5% for target speakers against impersonator's natural voice to an identification error rate of 22% for target speakers against impersonators' modified voice. These results show that the inclusion of prosodic and source-related features in the feature set for an automatic speaker recognition system requires careful consideration of the concomitant risk of impersonation, particulary by trained professional imitators. However, as the database is small, the results should be interpreted with caution.

5.3 Robustness analysis to converted voices

This section analyses the robustness of an automatic speaker recognition system against converted voices. The conversion system used to get such converted voices comes up from the improvement of a synthesis system based on the harmonic plus stochastic model (Erro and Moreno, 2005), which uses frames of fixed length, and where a conversion module has been implemented. The performance of the systems has been demonstrated to be notable, even when no training parallel corpus is available. This is partly due to the fact that the system takes advantage of the high flexibility of the harmonic plus stochastic model in order to minimise the errors derived of the signal reconstruction from their already modified parameters (Erro et al., 2007).

Next, the voice synthesis system (5.3.1) and the voice conversion method (5.3.2) are described. In order to analyse the robustness of an automatic speaker recognition system against converted voices (described in 5.3.3), the system is tested against both original and converted voices (5.3.4), so that the comparison will allow to see whether the performance gets worse by using voice conversion.

5.3.1 Voice synthesis system

The harmonic plus stochastic model (Stylianou, 1996) assumes that a speech signal can be represented as a sum of two components (Equation 5.1). The first one is the sum of harmonically related sinusoids with time-varying parameters that model the voiced part of the speech. This harmonic component is only present in the voiced segments and it can be represented in each analysis frame by the fundamental frequency and the amplitudes and phases of the harmonics.

The second one —the stochastic component— is a noise-like component, which is characterised by a particular power spectral density and models all non-sinusoidal components, both the unvoiced parts of the signal and the non-periodic phenomenon, like frication, breathing noise, etc. This component is represented in each frame by the coefficients of an all-pole filter.

$$s[n] = \sum_{j=1}^{J(n)} A_j[n] \cos(\theta_j[n]) + e[n] \, . \tag{5.1}$$

Analysis and reconstruction of the signal

Although both harmonic and stochastic components vary over time, they can be regarded as stable component within small intervals. In the current system, signals are analysed using equally-spaced frames of 10 ms. Given a speech frame k, the fundamental frequency $F0^{(k)}$ is estimated using the electrographic signal recorded simultaneously with the voice. Then, a binary voicing decision is taken; if the frame is detected as voiced, the amplitudes $\{A_j^{(k)}\}$ and phases $\{\varphi_j^{(k)}\}$ of all the harmonics below a cutoff frequency of 5 kHz are extracted (Depalle and Hélie, 1997).

The choice of a fixed cutoff frequency is adequate for voice conversion purposes, since the spectral envelopes are extracted from the harmonic component and they need to be parameterised. The harmonic part is interpolated over the different frames (McAulay and Quatieri, 1986) and subtracted from the original signal. The remaining part of the signal, which corresponds to the stochastic component, is analysed in each frame using the LPC technique.

The signal is reconstructed by overlapping and adding $2N$-length frames, where N is the distance between the analysis frame centres. Each synthetic frame contains the sum of the measured harmonics with constant amplitudes, frequencies and phases, and the stochastic contribution, generated by filtering white Gaussian noise with the measured LPC filters. A triangular window is used to overlap-add the frames in order to obtain the time-varying synthetic signal. Being k and j the frame and the harmonic number, respectively, the following expressions are used to reconstruct the signal:

$$s^{(k)}[n] = \sum_{j=1}^{J^{(k)}} A_j^{(k)} \cos\left(2\pi j \frac{f_0^{(k)}}{f_s} n + \varphi_j^{(k)}\right) + \sigma[n] * h_{LPC}^{(k)}[n] \, , \tag{5.2}$$

where $n = -N, ..., N-1$, and

$$s[kN+m] = \left(\frac{N-m}{N}\right) s^{(k)}[m] + \left(\frac{m}{N}\right) s^{(k+1)}[m-N] \, , \tag{5.3}$$

where m lies in the range $[0, N-1]$. The speech signal resynthesised from the measured parameters is almost indistinguishable from the original.

Prosodic modifications

Conversion of prosody —pitch evolution, speech rate, etc.— is out of the scope of this system. Only duration and fundamental frequency modifications by simple mean and variance adjustments are carried out, and they will be referred to as prosodic modifications.

Since fixed-length frames are being used, some techniques allowing the modification of the parameters in each signal frame in a simple way —without altering the phase coherence between frames— needed to be developed. To this end, new strategies to manipulate the phases were proposed, considering that the measured phases at a certain analysis frame are the sum of two components: a linear-in-frequency term and the phase contribution of the time-varying vocal tract.

Duration modification The duration modification can be carried out by altering the distance N between the synthesis point in Equation 5.3, in order to adapt the amplitude and fundamental frequency variations to the new time scale. Nevertheless, if the phases are kept unmodified at the centre of the frames, the waveform coherence between consecutive points is lost, causing artifacts and noisy pitch variations. Therefore, the change in N needs to be compensated with a phase manipulation, so that the waveform and pitch of the duration-modified signal are similar to the original.

This manipulation should only affect the linear-in-frequency phase term, keeping the phase contribution of the vocal tract unaltered in each frame without losing the coherence. Assuming that the vocal tract is not altered and that the fundamental frequency varies linearly between frames $k-1$ and k, the expected phase increment of the first harmonic between those points can be expressed by the following function:

$$\Psi(F_0^{(k-1)}, F_0^{(k)}, N) = (F_0^{(k-1)}, F_0^{(k)})\pi N \frac{1}{F_s} \ , \tag{5.4}$$

being F_s the sampling frequency. Replacing N by N', the following phase correction is applied:

$$\Delta\varphi_1^{(k)} = \Psi(F_0^{(k-1)}, F_0^{(k)}, N') - \Psi(F_0^{(k-1)}, F_0^{(k)}, N) \ , \tag{5.5}$$

$$\varphi'^{(k)}_1 = \varphi_1^{(k)} + j\sum_{q=1}^{k} \Delta\varphi_1^{(q)}, \quad j = 1...J^{(k)}, \quad \forall k \ , \tag{5.6}$$

being N' the new frame length and Ψ the function defined in 5.4. This correction compensates the modification of N without affecting the small local variations in the vocal tract phase response. Furthermore, the stochastic coefficients are not modified, and the modification factor can be time-varying.

Pitch modification In order to modify the pitch, the frequencies are multiplied by the desired factor and the harmonics are recalculated up to 5 kHz. The amplitudes of the new harmonics are obtained by a simple linear interpolation between the measured log-amplitudes in the analysis

(Banga et al., 2001), in order to keep the formant structure unaltered. A constant multiplicative factor is used to maintain the original energy of the harmonic component, despite the variation of the number of sinusoids. If the linear phase term is eliminated, the vocal tract phase response in the new harmonics $\theta'^{(k)}_j$ can be obtained by means of a linear interpolation of the real and imaginary parts of the complex amplitudes (Banga et al., 2001).

The linear term is eliminated as follows (Chazan et al., 2002):

$$\varphi^{(k)}_{VTj} = \varphi^{(k)}_j - j\alpha^{(k)} \tag{5.7}$$

$$\alpha^{(k)} = \arg\min \sum\nolimits_{j=0}^{J^{(k)}-1} \left| \sqrt{A^{(k)}_j} e^{i(\varphi^{(k)}_j - j\alpha^{(k)})} - \sqrt{A^{(k)}_{j+1}} e^{i(\varphi^{(k)}_{j+1} - (j+1)\alpha^{(k)})} \right|^2 , \tag{5.8}$$

taking $A^{(k)}_0 = A^{(k)}_1$ and $\varphi^{(k)}_0 = 0$. Once the vocal tract phase in the new harmonics is calculated, the linear term is replaced by using the same α value. In this case, the linear phase term should be also updated, since the period length is altered while N remains constant. The phase correction is given by (5.6) with the following increment:

$$\Delta\varphi^{(k)}_1 = \Psi(F'^{(k-1)}_0, F'^{(k)}_0, N) - \Psi(F^{(k-1)}_0, F^{(k)}_0, N) \; . \tag{5.9}$$

The stochastic coefficients are not modified. Time-varying modification factors can also be used with this method, and both modifications can be performed simultaneously.

Concatenation of units

Concatenative speech synthesis builds the synthetic utterances by concatenating different speech units selected from a recorded database. Given such database, the unit concatenation may follow particular specifications on duration, energy and pitch contour. Pertinent prosodic transformations are applied to each unit by modifying their amplitudes, frequencies, phases and LPC filter gains.

Since a pitch-asyncronous scheme is used, the phase coherence problem between units arises again. Once the prosody of the selected units is modified, the problem is avoided by making the linear phase term be continuous in the concatenation point. For each pair of adjacent units A and B, the limit points k_A (last frame of the last concatenated unit A) and k_B (first frame of the incoming unit B), and the corresponding $\alpha^{(k_A)}$ and $\alpha^{(k_B)}$ values according to Equation 5.8 are computed. Then, the phase correction is given by:

$$\varphi'^{(k)}_j = \varphi^{(k)}_j + j(-\alpha^{k_B} + \alpha^{k_A} + \Psi(F^{(k_A)}_0, F^{(k_B)}_0, N)), \quad j = 1...J^{(k)}, \quad k \geq k_B \; , \tag{5.10}$$

where Ψ is defined in 5.4. Finally, a smoothing technique is applied to the amplitudes of the harmonics near the concatenation point, so that the spectral discontinuities are minimised.

5.3.2 Voice conversion method

In order to optimise the transformation between two speakers, several speech aspects should be considered: intonation contours, particular lexical terms, etc. However, in this conversion system, only a spectral conversion is performed, together with a basic pitch transformation.

Pitch normalisation

Fundamental frequency is characterised by a log-normal distribution. Given the available data from each speaker, the mean and standard deviation of their $\log(F0)$ are estimated and the following normalisation is performed:

$$\log F_0^{(converted)} = \mu^{(target)} + \frac{\sigma^{(target)}}{\sigma^{(source)}}\Big(\log F_0^{(source)} - \mu^{(source)}\Big)\ . \tag{5.11}$$

Vocal tract conversion

Vocal tract can be modeled by using several types of parameters: amplitudes normalised to a fixed fundamental frequency, discrete cepstral coefficients, line spectral frequencies (LSF) coefficients, etc. LSF are the most commonly used coefficients in voice conversion, since the formant structure is properly modeled and the estimation error in one of these coefficients affects only a small part of the spectrum. Moreover, these coefficients also codify a minimum phase envelope, and the same codification is used in the stochastic component, which allows to relate both signal components.

In the current system, amplitudes are represented as LSF coefficients, using the all-pole filter that better fits the amplitudes of the harmonics calculated by means of the discrete all-pole modeling technique (El-Jaroudi and Makhoul, 1991). A 14-order filter is used for a sampling frequency of 16 kHz, since the contribution of superior orders is little relevant and they make the conversion more difficult. During the training phase, a GMM of $m = 8$ Gaussian components is trained from a set of phonetically aligned source-target vector pairs $v = [x^T y^T]^T$ (Kain and Macon, 1998), where vectors x and y contain 14 parameters corresponding to the source and target speaker, respectively.

The joint source-target GMM is represented by the weights α_i, the mean vectors μ_i and the covariance matrices Σ_i of each of the Gaussian components. Once the GMM has been trained, given a source LSF vector x, the probability of x belonging to the ith Gaussian component of the model $p_i(x)$ is given by:

$$p_i(x) = \frac{\alpha_i N(x, \mu_i^x, \Sigma_i^{xx})}{\sum_{j=1}^{m} \alpha_j N(x, \mu_j^x, \Sigma_j^{xx})}\ , \tag{5.12}$$

where $\mu_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix}$, $\Sigma_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix}$, and

$$N(\nu, \mu, \Sigma) = \tfrac{1}{(2\pi)^{p/2}} |\Sigma|^{-1/2} \exp\Big(-\tfrac{1}{2}(\nu - \mu)^T \Sigma^{-1} (\nu - \mu)\Big)\ .$$

The transformation function is then defined as:

$$F(x) = \sum_{i=1}^{m} p_i(x) \left[\mu_i^y + \Sigma_i^{yx} (\Sigma_i^{xx})^{-1} (x - \mu_i^x) \right] . \tag{5.13}$$

Phase envelope estimation

In order to obtain a high-quality synthetic voice, the temporal variations in vocal tract magnitude must be related to their corresponding phase variations, maintaining, at the same time, the phase coherence between frames. In order to satisfy both conditions, a linear phase term is calculated recursively:

$$\varphi_i^{(k)} = \varphi_i^{(k-1)} + j\Psi\left(f_0^{(k-1)}, f_0^{(k),N} \right) . \tag{5.14}$$

At the beginning of each voiced region, phases are initialised to zero. In this first step, no considerable phase discontinuities are caused, but as no vocal tract phase variations exist between frames, the synthesised signal could contain an annoying metallic noise. Therefore, an additional term is included:

$$\varphi'^{(k)}_i = \varphi_i^{(k)} + \arg(1/A(f_j^{(k)})) , \tag{5.15}$$

being $1/A(f_j^{(k)})$ the all-pole filter converted from the harmonic amplitudes. In fact, this term does not represent the vocal tract phase envelope, but it provides coherent phase variations with amplitudes, achieving a rather acceptable quality.

Stochastic component prediction

In order to predict the stochastic component of the target speaker from the LSF representation of its harmonic component, a new function is designed. The stochastic LPC coefficients associated to the training LSF vectors y are also translated into LSF vectors y_{st}, and matrices Σ_i and vectors η_i are found so that the following prediction function is optimised in the training phase:

$$y_{st} = \sum_{i=1}^{m} p_i(y)[\eta_i + \Gamma_i(\Sigma_i^{yy})^{-1}(y - \mu_i^y)] , \tag{5.16}$$

where y_{st} is the stochastic LSF vector, and μ_i^y and Σ_i^{yy} are used in Equation 5.12 to obtain $\pi(y)$. In the conversion phase, the stochastic component of the converted frame is predicted by applying Equation 5.16, where y is replaced by the converted LSF vector $F(x)$ previously computed in 5.13.

No transformation has been performed on the unvoiced frames, since their conversion does not lead to considerable improvements in the perceptual aspect and, besides, it can cause loss of quality.

5.3.3 Voice conversion database

The characteristics of the database used for voice conversion are summarised in Table 5.10. This material was made available by UPC for the evaluation campaigns of the TC-STAR project (Bonafonte et al., 2006). The voice conversion corpora contain around 200 sentences in Spanish and 170 in English —although only the Spanish ones were used in these experiments— uttered by four different professional bilingual speakers, two males and two females. The average duration of the sentences was four seconds, so that about 10-15 minutes of audio were available for each speaker and language.

Number of voices	2 male voices: M1, M2 2 female voices: F1, F2
Language	Spanish and English (bilingual speakers)
Amount of data	Spanish: ∼200 sentences, ∼4 seconds/sentence English: ∼170 sentences, ∼4 seconds/sentence
Type of utterances	Mimic parallel sentences

Table 5.10: General characteristics of the voice conversion database.

The sentences uttered by the speakers are exactly the same, so that parallel training corpora can be used for training voice conversion functions. In addition, the sentences were recorded as mimic sentences. This means that there were no significant prosodic differences between speakers, since they all were asked to imitate the same prerecorded pattern with neutral speaking style for each of the sentences.

5.3.4 Speaker identification using original and converted voices

First of all, the original data set consisting of all four voices described in the previous section was divided in three sets of sentences. The first set was set aside to train the transformation function of the conversion system, and the second and third set of sentences were used to train and test the automatic recognition system, respectively.

Each of the four original voices was converted to the rest of the voices. Since there are twelve pairs of source-target voices, a set of twelve converted voices was obtained: four sets corresponding to intra-gender conversions (*female to female* and *male to male* conversions), and eight sets corresponding to cross-gender conversion (*female to male* and *male to female* conversions). Each set of converted voices consisted of 100 sentences.

The transformation function for the conversion system was trained using 10, 30 and 80 pairs of source-target sentences. Other 10 original sentences were used to train each of the four speaker models of the recognition system, and 100 more original sentences, together with the converted sentences, were used for testing.

The recognition system utilised in the identification experiments was a conventional 32-component GMM system, using short-term feature vectors consisting of 20 MFCC with a frame size of 24 ms and a shift of 8 ms. The corresponding delta and acceleration coefficients were also included.

In order to test the performance of the recognition system, a preliminary experiment was conducted by using only the original voices. Table 5.11 shows the corresponding identification matrix, where 100 sentences of each original voice were identified from the closed set of four speaker models. Since it was a rather simple experiment that used a low amount of speakers, a high performance was obtained, leading to a percent identification of 100% in three of the four voices. Only one of the males (M1) was once confused with the other male (M2), which suggests —given the high performance of the system— that both male voices are characterised by a significant degree of similarity.

	F1	**F2**	**M1**	**M2**
F1	100	0	0	0
F2	0	100	0	0
M1	0	0	99	1
M2	0	0	0	100

Table 5.11: Identification matrix for two male (M) and two female (F) original voices.

The identification experiments were conducted by testing the intra-gender and cross-gender converted voices. The system tried to identify 100 sentences of each converted voice again from the closed set of four speaker models. Moreover, three sets of converted voices were identified, according to the sentences used in training the transformation function (10, 30 or 80), in order to see how the amount of training data used in the conversion phase influenced the performance of the recognition system.

Source	**Target voice**			
voice	**F1**	**F2**	**M1**	**M2**
F1	-	-	0	0
F2	0	-	0	0
M1	0	0	-	0
M2	0	16	93	-

(a) *Source* identification.

Source	**Target voice**			
voice	**F1**	**F2**	**M1**	**M2**
F1	-	-	46	100
F2	100	-	98	100
M1	100	98	-	100
M2	100	84	7	-

(b) *Target* identification.

Source	**Target voice**			
voice	**F1**	**F2**	**M1**	**M2**
F1	-	-	54	0
F2	0	-	2	0
M1	0	2	-	0
M2	0	0	0	-

(c) *Other* identification.

Table 5.12: *Source* (a), *target* (b) and *other* (c) identifications using 10 sentences in training the transformation function.

Tables 5.12, 5.13 and 5.14 show the identification results corresponding to the number of

Source	Target voice			
voice	F1	F2	M1	M2
F1	-	0	0	0
F2	0	-	0	0
M1	0	0	-	0
M2	0	9	92	-

(a) *Source* identification.

Source	Target voice			
voice	F1	F2	M1	M2
F1	-	99	43	100
F2	100	-	95	100
M1	100	98	-	100
M2	100	91	8	-

(b) *Target* identification.

Source	Target voice			
voice	F1	F2	M1	M2
F1	-	1	57	0
F2	0	-	5	0
M1	0	2	-	0
M2	0	0	0	-

(c) *Other* identification.

Table 5.13: *Source* (a), *target* (b) and *other* (c) identifications using 30 sentences in training the transformation function.

sentences used to train the transformation function: 10, 30 and 80, respectively. (The converted F1_to_F2 voices by using 10 training sentences were damaged and not available at the time of doing these experiments). In each table, three types of identification are distinguished:

(a) source: where the converted voice was identified as its corresponding source speaker,

(b) target: where the converted voice was identified as its corresponding target speaker, and

(c) other: where the converted voice was identified as a speaker other than the corresponding source and target speakers.

The identification results corresponding to 30 training sentences are also plotted in Figure 5.3, in which the identification types are also represented by different colours: green, yellow and red for *source*, *target* and *other* identifications, respectively.

Regarding intra-gender identification, the results show that most of the converted voices were identified as their target voices, so that the recognition system failed in identifying the converted voice as the real source voice. Nevertheless, there is one case in which the performance of the system was better —or, in other words, where the voice conversion was not so successful—; this is the conversion of the second male to the first male (M2_to_M1), where most of the speakers were identified as the original source voice (M2) instead of as the target voice (M1). This could probably be explained by the fact that speaker M2 may be highly characterised by his unvoiced

Source voice	**Target voice**			
	F1	**F2**	**M1**	**M2**
F1	-	0	0	0
F2	0	-	0	0
M1	0	0	-	0
M2	0	5	72	-

(a) *Source* identification.

Source voice	**Target voice**			
	F1	**F2**	**M1**	**M2**
F1	-	100	87	100
F2	100	-	100	100
M1	100	99	-	100
M2	100	95	28	-

(b) *Target* identification.

Source voice	**Target voice**			
	F1	**F2**	**M1**	**M2**
F1	-	0	13	0
F2	0	-	0	0
M1	0	1	-	0
M2	0	0	0	-

(c) *Other* identification.

Table 5.14: *Source* (a), *target* (b) and *other* (c) identifications using 80 sentences in training the transformation function.

segments, and since these are not converted by the system, this unvoiced characteristics still remain in the converted M2_to_M1 voice. However, the identification as the source voice — which will be referred to as *correct identification* by convention— decreases as the amount of conversion training data increases.

It seems thus that the conversion system has difficulties in converting M2 to M1, which could be explained by the fact (seen in Table 5.11) that both M1 and M2 seem to be similar. However, the reverse phenomenon (M1_to_M2 identified as M1) is not observed in these experiments. Moreover, since the converted F1_to_F2 voice is strangely identified as the male speaker M2 in Table 5.11, it seems that the recognition system has a slight tendency to identify any speaker as M2.

On the other hand, half of the eight sets of cross-gender converted voices lead to a *miss identification* and *correct conversion* equaling 100%; i.e. not only were the converted speakers not identified as the corresponding source speaker (*miss identification*) but they were all identified as the corresponding target speaker (*correct conversion*).

The other half of the cross-gender conversions were not completely recognised as their corresponding target voices. These are those conversions trying to convert a female speaker to M1 and a male speaker to F2. All the errors are a miss conversion to speaker M2, except in the conversion M2_to_F2, where the errors can be seen, in fact, as a correct identification of the speaker M2. The worse results are found in the F1_to_M1 conversion, where the tendency of the system to identify speakers as if they were speaker M2 is summed to the hypothetic similarity between

M1 and M2 seen in Table 5.11. In all cases, however, an increase of the correct conversion is observed when the transformation function is trained using 80 sentences.

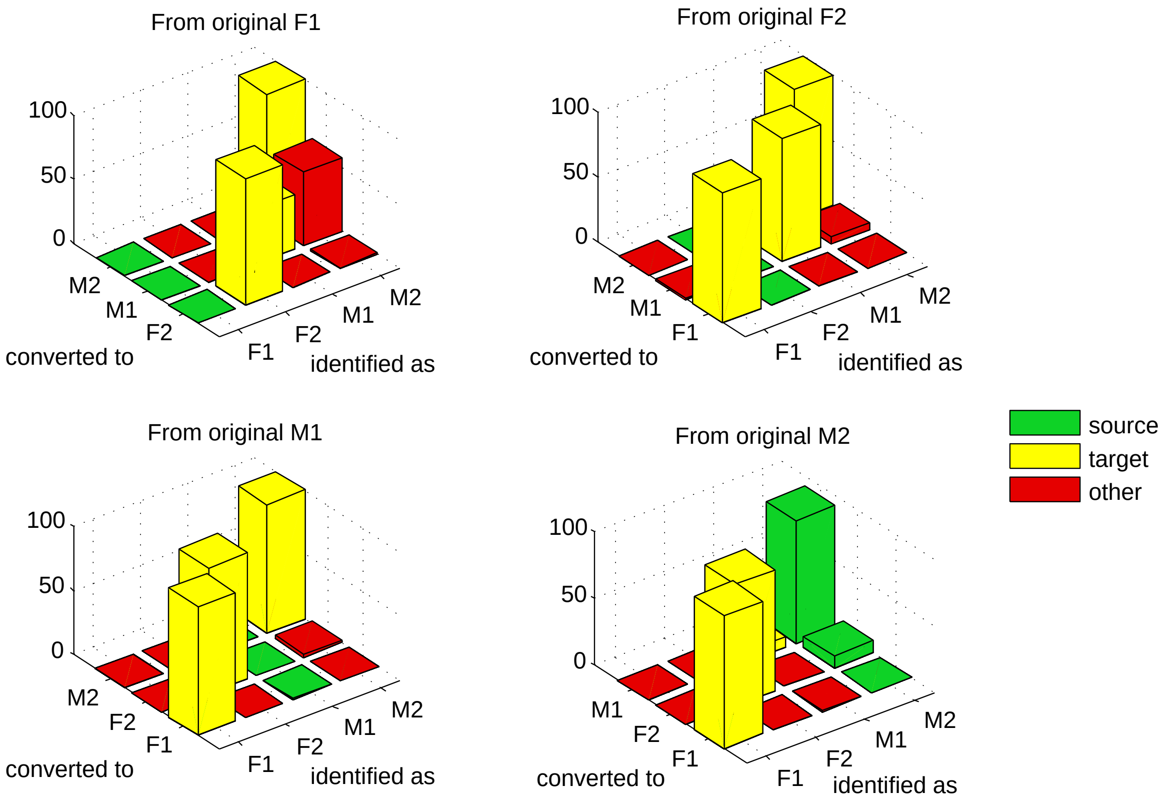


Figure 5.3: Identification of each converted voice using 30 sentences in the transformation function. Green, yellow and red bars indicate *source*, *target* and *other* identification, respectively.

Summarising, Table 5.15 shows the types of identification generated by both intra-gender and cross-gender conversions using 30 training sentences, which are also plotted in Figure 5.4. In general terms, intra-gender conversion tends to be identified as its corresponding source speaker in a higher degree than cross-gender conversion. On the other hand, cross-gender conversion tends to be more *successful* —speaking in conversion terms— than the intra-gender one, since the percentage of target identification is greater. Nevertheless, cross-gender conversion also leads to a higher percentage of *other* identification; ie. an erroneous conversion in which the converted voice is not identified as either of the source and target speakers.

Conversion type	Source	Target	Other
Intra-gender	23.0%	76.7%	0.3%
Cross-gender	1.1%	90.9%	8.0%

Table 5.15: Identification in percent of intra-gender and cross-gender conversions depending on the type of identification generated (*source*, *target* and *other*), where the transformation function has been trained using 30 sentences.

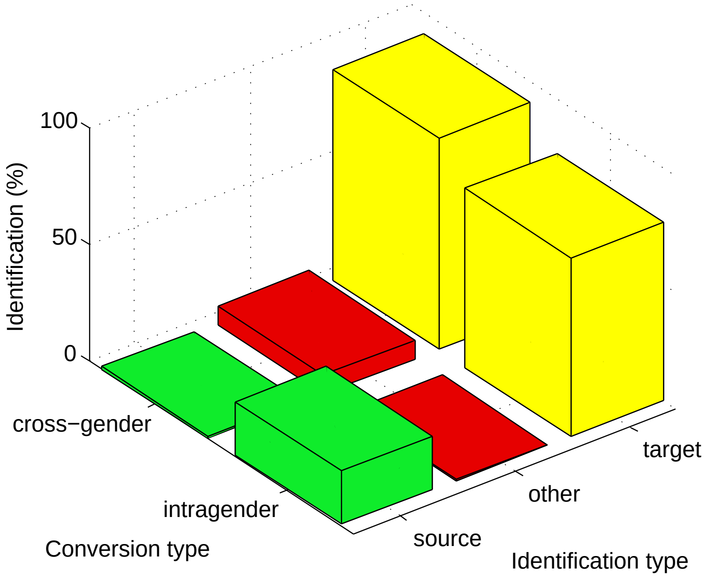


Figure 5.4: Identification of intra-gender and cross-gender conversions using 30 training sentences depending on the type of identification generated (*source*, *target* and *other*).

5.4 Conclusions

One of the handicaps in trying to know how recognition systems react to imitated and converted voices is the lack of databases consisting of this sort of disguised voices, specially when dealing with human imitations. Setting aside the lack of a consistent amount of training data, the experiments presented in this chapter point out to some interesting results; in section 5.1, for example, showed that dialect imitation can confuse both the human listener and speaker recognition systems —yet in different ways— and that high-quality dialect disguise is a topic that warrants forensic linguistic consideration. Other experiments conducted in section 5.2, where professional imitators attempted to mimic a target voice, showed that the identification error rate for most of the selected prosodic and source-related features increased when testing the mimicking voices with respect to the natural voices.

On the other hand, some experiments performed in section 5.3 tried to analyse the behaviour of an automatic speaker recognition system in front of automatic converted voices. They showed that most of the converted voices were identified as their corresponding target speaker; however, they failed sometimes to deceive the system and the source voice was recognised, especially in the intra-gender conversions, which leads to think that the recognition system may be more robust to these kind of conversions than the cross-gender ones. The results also pointed out that some voices are more difficult to convert than others, and that the correct identification decreases as the amount of conversion training data increases.

Nevertheless —as it was said above— the amount of training data is small enough to interpret the results with extreme caution. The main objective of this chapter was not to reinforce existing experiments and strengthen new conclusions, but to propose some research guidelines in order to investigate in more detail the influence of disguised voices in the speaker recognition task.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Several works have demonstrated that the use of prosodic information helps to improve recognition systems based solely on spectral parameters. This fact was also corroborated in the experiments performed in chapter 3 of the current thesis, where a preliminary speaker verification system based on prosodic features was built in order to improve a voice spectrum-based verification system over the conversational Switchboard-I database. In this thesis, the performance of the spectral system was considerably improved by using those prosodic features related to segment duration and fundamental frequency; however, the information contained in pauses did not result in an improvement either of the spectral and the rest of prosodic features.

The experiments in chapter 3 also showed the importance in which the way the scores are fused and the distance used in the decision classifier: Kullback-Leibler divergence outperformed, in almost all cases, Euclidean and Mahalanobis distances. However, when choosing different normalisation techniques, no big differences appeared between both min-max and z-score normalisations, except where pauses and spectral information were involved; in this case, the use of min-max normalisation resulted in a better performance. In respect to fusion techniques utilised in the experiments, matcher weighting outperformed, in most cases, simple sum, since the former takes advantage of the performance (EER) of each individual feature.

Also in chapter 3, additional acoustic features —namely jitter and shimmer— were used in order to improve a speaker verification system based on prosodic and spectral parameters. Jitter and shimmer measurements analyse the perturbation of fundamental frequency and waveform amplitude, respectively. The results showed that jitter and shimmer can be used to provide complementary information to both spectral and prosodic systems, and that the absolute measurements of both jitter and shimmer parameters seem to be more speaker discriminant than their correspondent relative measurements.

The prosodic system proposed in chapter 3 was also used in chapter 4, where prosodic features were added in a multimodal environment, improving the performance of a bimodal system based on facial and spectral information. The results varied largely depending on the fusion technique utilised. In this thesis, for instance, the use of support vector machines (a non-linear classifier) outperformed the overall fusion results obtained with matcher weighting technique. In addition, the experiments showed that the overall results can be improved by applying a histogram equalisation as a normalisation technique.

The experiments conducted in chapter 4 also showed the relevance in the way how the scores are fused for the performance of the multimodal system. Multimodal fusion was conducted in different strategies depending on the number steps involved in the fusion process: one, two, and three. In one-step fusion, all the features involved were fused at once, while in two- and three-step fusion the speech features were previously fused at once or in two steps before being combined with facial scores. When using z-score normalisation and matcher weighting technique in the trimodal system, the best results were obtained in one-step fusion strategy. On the contrary, support vector machines performed better in one of the configurations of the two-step fusion or in the three-step fusion, depending on the dimension of the prosodic feature vector used in the verification task. Moreover, the experiments performed over the extensive database allowed to observe that a previous fusion of the speech information —spectral and prosodic scores— did not contribute to improve the system.

The last chapter of the thesis deals with the robustness of speaker recognition to imitated and converted voices, where several experiments were designed in order to test the vulnerability of the speaker recognition task in different imitation environments. The first experiment, for instance, showed that dialect imitation can confuse both the human listener and speaker recognition systems. In a second experiment, professional imitators attempted to mimic a well-known target voice. As a result, the identification error rate for most of the selected prosodic and source-related features increased when testing the mimicking voices with respect to the natural voices. Finally, some experiments tried to analyse the behaviour of an automatic speaker recognition system against converted voices. Most of the converted voices were identified as their corresponding target speakers; however, in some of the experiments, they failed to deceive the system and the source voice was recognised. In others, the systems output as decision a speaker that was neither the source not the target. The results point out that some voices are more difficult to convert than others, and that using different spectral coefficients in the conversion and test phase may influence the recognition results.

6.2 Future work

It has been seen that additional linguistic levels help to improve the performance of the traditional systems using filter parameters. In the current thesis, prosodic information has been used to this end, and other recent works have also added other linguistic sources like phonetic information and lexical characteristics of the speakers, among others.

The way how human recognise others from voice alone turned out to be an important clue in defining these new information sources. Prosody, phonetic and lexical characteristic are maybe the most relevant ones. However, listeners also notice other voice singularities such as voice quality; breathy, rough and hoarse voices, for instance, are manifested in acoustic characteristics like jitter and shimmer, which have also been used in the current thesis.

There may be still more features that could be of importance in the speaker recognition task. Future work in this field might try to capture new characteristics —used or not used by human listeners— that convey useful speaker information, and find new and appropriate modeling techniques for this additional sources, since linguistic modeling is still a difficult and an in-development task.

One of the handicaps in using prosodic and other linguistic information is the need of large training data to guarantee a reasonable performance of the system. In addition, some modeling techniques make necessary a previous labeling of the database. These requirements hinder the use of these sources in real-time applications, which undergo an increasing demand. Therefore, future research should also be focused on improving the linguistic-based systems, making them more easy-to-use.

Moreover, linguistic information should also be added to multimodal biometric systems. Since the current study has shown that prosody helps to improve a spectral- and facial-based biometric recognition, there is reason to believe that such improvement could be extended to other biometric modalities. In this sense, research should also point to the development of normalisation and fusion techniques and strategies, trying to find the best methods depending on the biometrics used and the correlation between the involved modalities.

It has also been pointed out that voice imitation, in all its possible forms, is a potential threat to automatic recognition systems. One of the handicaps in trying to know how recognition systems react to imitated and converted voices is the lack of databases consisting of this sort of disguised voices, especially when dealing with human imitations. One of the main research objectives in this field should be the development of extensive databases consisting of human disguised voices. In respect to robustness to voice conversion, the research lines should focus on the influence of different conversion techniques and converted features on the vulnerability of automatic recognition systems. In parallel, the analysis of such system vulnerability should be carried out by using different parameters and recognition techniques, and it should also be extended to systems based on additional features different to the traditional ones.

Bibliography

A. Abad, C. Nadeu, J. Hernando, and J. Padrell. Jacobian adaptation based on the frequency-filtered spectral energies. In *Proceedings of the Eurospeech*, pages 1621–1624, Geneva, Switzerland, 2003.

A. Adami. Prosodic modeling for speaker recognition based on sub-band energy temporal trajectories. In *Proceedings of the ICASSP*, volume 1, pages 189–192, 2005.

A. Adami and H. Hermansky. Segmentation of speech for speaker and language recognition. In *Proceedings of the Eurospeech*, Geneva, Switzerland, 2003.

A. Adami, R. Mihaescu, D.A. Reynolds, and J.J. Godfrey. Modeling prosodic dynamics for speaker recognition. In *Proceedings of the ICASSP*, volume I, 2003.

A.G. Adami. Modeling prosodic differences for speaker recognition. *Speech Communication*, 49 (4):277–291, April 2007.

A. Alexander, D. Dessimoz, F. Botti, and D. Andrzej. Aural and automatic forensic speaker recognition in mismatched conditions. *The International Journal of Speech, Language and the Law*, 12(2):214–234, 2005.

W. Andrews, M.A. Kohler, J. Campbell, J. Godfrey, and J. Hernández-Cordero. Gender-dependent phonetic refraction for speaker recognition. In *Proceedings of the ICASSP*, volume I, pages 149–152, 2002.

W.D. Andrews, M.A. Kohler, and J.P. Campbell. Phonetic speaker recognition. In *Proceedings of the Eurospeech*, pages 2517–2520, Aalborg, Denmark, 2001.

M. Arcienega and A. Drygaljo. Pitch-dependent GMMs for text-independent speaker recognition systems. In *Proceedings of the Eurospeech*, pages 2821–2825, Aalborg, Denmark, September 2001.

B.S. Atal. Automatic speaker recognition based on pitch contours. *Journal of the Acoustical Society of America*, 52:1687–1697, 1972.

J.E. Atkinson. Correlation analysis of the physiological factors controling fundamental voice frequency. *Journal of the Acoustical Society of America*, 63(1):211–222, 1978.

R. Balchandran and R. Mammone. Non parametric estimation and correction of non linear distortion in speech systems. In *Proceedings of the ICASSP*, 1998.

J.R. Baldwin. Phonetics and speaker identification. *Medicine, Science and the Law*, 19:231–232, 1979.

J.R. Baldwin and P. French. *Forensic Phonetics.* Pinter Publishers, London, 1990.

E.R. Banga, C. García-Mateo, and X. Fernández-Salgado. Concatenative text-to-speech synthesis based on sinousoidal modelling. In *Improvements in Speech Synthesis*, pages 39–51. John Wiley and Sons, Ltd., 2001.

K. Bartkova, D. Le-Gac, D. Charlet, and D. Jouvet. Prosodic parameter for speaker identification. In *Proceedings of the ICSLP*, Colorado, 2002.

M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18:349–369, 1989.

M. Behlau, G. Madazio, D. Feijó, and P. Pontes. *Avaliação da Voz*, volume I, chapter 3, pages 86–180. Revinter, 2001.

M. Behlau and P. Pontes. *Avaliação e Tratamento das Disfonias.* Lovise, São Paulo, 1995.

P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

R.E. Bellman. *Adaptive Control Processes.* Princeton University Press, Princeton, NJ, 1961.

Y. Bennani and P. Gallinari. On the use of TDNN-extracted features information in talker identification. In *Proceedings of the ICASSP*, pages 385–388, 1991.

F. Bimbot, J. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D.A. Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 4:430–451, 2004.

P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *IFA Proceedings*, 17:97–110, 1993.

P. Boersma and D. Weenink. *Praat: Doing phonetics by computer (Version 4.5.16, Computer program, retrieved from from http://www.praat.org).* Institute of Phonetic Sciences, University of Amsterdam, Amsterdam, 1992.

R.M. Bolle, J.H. Connell, S. Pankanti, N.K. Ratha, and A.W. Senior. *Guide to Biometrics.* Springer, New York, 2004.

R.H. Bolt, F.S. Cooper, E.E. Jr. David, P.B. Denes, J.M. Pickett, and K.N. Stevens. Identification of a speaker by speech spectrograms. *Science*, 166, October 1969.

A. Bonafonte, H. Höge, I. Kiss, A. Moreno, U. Ziegenhain, H. van den Heuvel, H.U. Hain, X.S. Wang, and M.N. Garcia. TC-STAR: Specifications of language resources and evaluation for speech synthesis. In *Proceedings of the International Conference on Language Resources and Evaluation*, Genoa, Italy, 2006.

L. Boves and H. Strik. The fundamental frequency-subglottal pressure ratio in speech. *Journal of the Acoustical Society of America*, 84(1):82, 1988.

J.D. Brand, J.S. Mason, and S. Colomb. Visual speech: a physiological or behavioural biometric? In *AVBPA*, volume 2091 of *Lecture Notes in Computer Science*, 2001.

R. Brunelli and D. Falavigna. Person identification using multiple cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):955–966, 1995.

C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge discovery*, 2:121–167, 1998.

J.P. Campbell. Speaker recognition: A tutorial. In *Proceedings of the IEEE*, volume 85, pages 1437–1462, 1997.

J.P. Campbell, D.A. Reynolds, and R.B. Dunn. Fusing high- and low-level features for speaker recognition. In *Proceedings of the Eurospeech*, 2003.

M.J. Carey, E.S. Parris, H. Lloyd-Thomas, and S. Bennett. Robust prosodic features for speaker identification. In *Proceedings of the ICSLP*, volume 3, pages 1800–1803, Philadelphia, 1996.

K. Chang, K.W. Bowyer, S. Sarkar, and B. Victor. Comparison and combination of ear and face images in appearance-based biometrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1160–1165, 2003.

D. Chazan, R. Hoory, Z. Konz, D. Silberstein, and A. Sorin. Reducing the footprint of the IBM trainable speech synthesis system. In *Proceedings of the ICSLP*, Denver, Colorado, 2002.

R.F. Coleman, J. Mabis, and J. Hinson. Fundamental frequency-sound pressure level profiles of adult male and female voices. *Journal of Speech and Hearing Research*, 20:197–204, 1977.

R. Collier. Physiological correlates of intonation patterns. *Journal of the Acoustical Society of America*, 58(1):249–255, 1975.

R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18 (1):32–80, 2001.

N. Cristianini and J. Shawe-Taylor. *An introduction to Support Vector Machines (and other kernel-based learning methods).* Cambridge University Press, 2000.

P. Cunningham and S.J. Delany. k-Nearest Neighbour Classifiers. Technical report UCD-CSI-2007-4, University College Dublin, Dublin Institute of Technology, Dublin, 2007.

S.B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 28:357–366, 1980.

A. de la Torre, A.M. Peinado, J.C. Segura, J.L., Pérez-Córdoba, M.C. Benítez, and A.J. Rubio. Histogram equalization of speech representation for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(3):355–366, 2005.

D.D. Deliyski. Acoustic model and evaluation of pathological voice production. In *Proceedings of the Eurospeech*, pages 1969–1972, Berlin, Germany, 1993.

V. Dellwo, M. Huckvale, and M. Ashby. How is individuality expressed in voice? An introduction to speech production and description for speaker classification. In Christian Müller, editor, *Speaker Classification I*, Lecture Notes in Computer Science, pages 1–20. Springer, Berlin, 2007.

P. Depalle and T. Hélie. Extraction of spectral peak parameters using a short-time Fourier transform modeling and no sidelobe windows. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, 1997.

G. Doddington. A method for speaker verification. *Journal of the Acoustical Society of America*, 49(1):139, 1971.

G. Doddington. Speaker recognition - identifying people by their voices. *Proceedings of the IEEE*, 73(11):1651–1663, November 1985.

G. Doddington. Some experiments on idiolectal differences among speakers. Motivating material for NIST 2001 Extended Training Task. Available in: http://www.nist.gov/speech/tests/sre/2001/index.html, 2000.

G. Doddington. Speaker recognition based on idiolectal differences between speakers. In *Proceedings of the Eurospeech*, volume 4, pages 2521–2524, 2001.

G. Doddington, M.A. Przybocki, A.F. Martin, and D.A. Reynolds. The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective. *Speech Communication*, 31: 225–254, 2000.

E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40:33–60, 2003.

R.O. Duda. *Pattern Classification.* John Wiley & Sons, Inc., 2001.

H. Duxans. *Voice Conversion applied to Text-to-Speech systems.* PhD thesis, Universitat Politcnica de Catalunya, Barcelona, 2006.

H. Duxans, D. Erro, J. Pérez, F. Diego, A. Bonafonte, and A. Moreno. Voice conversion of non-aligned data using unit selection. In *Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, 2006.

A. El-Jaroudi and J. Makhoul. Discrete all-pole modeling. *IEEE Transactions on Signal Processing*, 1991.

E.J. Eriksson, F. Schaeffler, M. Sjöström, K.P.H. Sullivan, and E. Zetterholm. On the perceptual dominance of dialect (Manuscript). *Perception & Psychophysics*, 2007.

D. Erro and A. Moreno. A pitch-asynchronous simple method for speech synthesis by diphone concatenation using the deterministic plus stochastic model. In *Proceedings of the SPECOM*, Patras, Greece, 2005.

D. Erro and A. Moreno. Sistema de síntesis armónico/estocástico en modo pitch-asíncrono aplicado a conversión de voz. In *Proceedings of the IV Jornadas en Tecnología de Habla*, Zaragoza, 2006.

D. Erro, A. Moreno, and A. Bonafonte. Flexible harmonic/stochastic speech synthesis. In *Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW6)*, Bonn, Germany, August 2007.

G. Fant. *Acoustic Theory of Speech Production.* Mouton and Co., The Hague, Netherlands, 1960.

G. Fant. Preliminaries to analysis of the human voice. Technical report, Royal Institute of Technology, Department of Speech, Music & Hearing, Stockholm, Sweden, 1982.

M. Farrús, P. Ejarque, A. Temko, and J. Hernando. Histogram equalization in SVM multimodal person verification. In *Proceedings of the IEEE International Conference on Biometrics*, volume 4642 of *Lecture Notes in Computer Science*, pages 819–827, August 2007.

M. Farrús, E. Eriksson, K.P.H. Sullivan, and J. Hernando. Dialect imitations in speaker recognition. In *2nd European IAFL Conference on Forensic Linguistics, Language and the Law*, pages 347–353, Barcelona, September 2006a.

M. Farrús, A. Garde, P. Ejarque, J. Luque, and J. Hernando. On the fusion of prosody, voice spectrum and face features for multimodal person verification. In *Proceedings of the ICSLP*, pages 2106–2109, Pittsburgh, 2006b.

J. Flanagan. *Speech Analysis, Synthesis and Perception*. Springer, Berlin-Heidelberg-New York, 1972.

N.A. Fox, R. Gross, P. Chazal, J.F. Cohn, and R.B. Reilly. Person identification using automatic integration of speech, lip and face experts. In *Proceedings of the ACM SIGMM 2003 Multimedia Biometrics Methods and Applications Workshop*, pages 25–32, Berkeley, CA, 2003.

P. French. An overview of forensic phonetics with particular reference to speaker identification. *Forensic Linguistics*, 1(2):169–181, 1994.

S. Furui. Speaker independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(1):52–59, February 1986.

S. Furui. An overview of speaker recognition technology. In *Automatic Speech and Speaker Recognition: Advanced Topics*, pages 31–56. Academic Publishers, Norwell, MA, 1996.

D. García-Romero, J. Fiérrez-Aguilar, J. Ortega-García, and L. González-Rodríguez. Support vector machine fusion of idiolectal and acoustic speaker information in Spanish conversational speech. In *Proceedings of the ICASSP*, volume 2, pages 229–232, 2003.

H. Gish and M. Schmidt. Text-independent speaker identification. *IEEE Signal Processing Magazine*, 11(4):18–32, October 1994.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the ICASSP*, pages 517–520, 1990.

N. Halloran. The acquisition of a stage dialect. Master's thesis, Portland State University, 2003.

M.A. Hearst. Trends and controversies: Support vector machines. *IEEE Intelligent Systems*, 13: 18–28, 1998.

M.H.L. Hecker, K. Stevens, G. von Bismark, and C. Williams. Manifestations of task-induced stress in the acoustic speech signal. *Journal of the Acoustical Society of America*, 44:993–1001, 1968.

J. Hernando. CDHMM speaker recognition by means of frequency filtering of filter-bank energies. In *Proceedings of the Eurospeech*, 1997.

J. Hernando, M. Farrús, A. Garde, P. Ejarque, and J. Luque. Person verification by fusion of prosodic, voice spectral and facial parameters. In *Proceedings of the International Conference on Security and Cryptography (SECRYPT)*, pages 17–23, Setúbal, Portugal, 2006.

A. Higgins, L. Bahler, and J. Porter. Voice identification using nearest neighbor distance measure. In *Proceedings of the ICASSP*, pages 375–378, 1993.

A. Higgins, L. Bahler, and J. Porter. Voice identification using nonparametric density matching. In *Automatic Speech and Speaker Recognition: Advanced Topics*, pages 211–232. Kluwer Academic Publishers, Norwell, MA, 1996.

F. Hilger and H. Ney. Quantile based histogram equalization for noise robust speech recognition. In *Proceedings of the Eurospeech*, pages 1135–1138, Aalborg, Denmark, 2001.

H. Hollien. *Forensic Voice Identification.* Academic Press, London, 2002.

H. Hollien and W. Majewski. Speaker identification by long-term spectra under normal and distorted speech conditions. *Journal of the Acoustical Society of America*, 62(4):975–980, 1977.

J.P. Hosom, A.B. Kain, T. Mishra, J.P.H. van Santen, M. Fried-Oken, and J. Staehely. Intelligibility of modifications to dysarthric speech. In *Proceedings of the ICASSP*, pages 924–927, 2003.

P.J. Huber. *Robust Statistics.* Wiley, New York, 1981.

M. Indovina, U. Uludag, R. Snelik, A. Mink, and A. Jain. Multimodal biometric authentication methods: A COTS approach. In *Proceedings of the Workshop on Multimodal User Authentication*, pages 99–106, Santa Barbara, CA, 2003.

S. Jaeger, H. Ma, and D. Doermann. Identifying script on word-level with informational confidence. In *Proceedings of the Eighth International Conference on Document Analysis and Recognition*, volume 1, pages 416–420, Seoul, Korea, August 2005.

A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 2005.

A.K. Jain and A. Ross. Learning user-specific parameters in a multibiometric system. In *Proceedings of the International Conference on Image Processing*, pages 57–60, New York, 2002.

A.K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics*, 14(1):4–20, January 1994.

Q. Jin, J. Navrátil, D.A. Reynolds, J.P. Campbell, W.A. Andrews, and J.S. Abramson. Combining cross-stream and time dimensions in phonetic speaker recognition. In *Proceedings of the ICASSP*, 2003.

A. Kain and M. Macon. Spectral voice conversion for text-to-speech synthesis. In *Proceedings of the ICASSP*, 1998.

S. Kajarekar, L. Ferrer, A. Venkataraman, K. Sönmez, E. Shriberg, A. Stolcke, H. Bratt, and R.R. Grade. Speaker recognition using prosodic and lexical features. In *Proceedings of the IEEE Speech Recognition and Understanding Workshop*, pages 19–24, 2003.

R.D. Kent and C. Read. *The Acoustic Analysis of Speech.* Whurr Publishers - Singular Publishing Group, London - San Diego, 1992.

L.G. Kersta. Voiceprint identification. *Nature*, 196(4861):1253–1257, December 1962.

R.L. Klevans and R.D. Rodman. *Voice Recognition.* Artech House, Inc., Boston, 1997.

D. Klusácec, J. Navrátil, D.A. Reynolds, and J.P. Campbell. Conditional pronunciation modeling in speaker recognition. In *Proceedings of the ICASSP*, 2003.

H. Kou and G. Gardarin. Study of category score algorithms for k-NN classifier. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 393–394, Tampere, Finland, 2002. ACM Press, New York, NY, USA.

J. Kreiman and B.R. Gerrat. Perception of aperiodicity in pathological voice. *Journal of the Acoustical Society of America*, 117(4):2201–2211, 2005.

A. Kumar, D.C.M. Wong, H.C. Shen, and A.K. Jain. Personal verification using palmprint and hand geometry biometric. In *Proceedings of the Fourth International Conference on Audio- and Video-based Biometric Person Authentication*, pages 668–678, Guildford, UK, 2003.

H.J. Künzel. *Sprecherkennung: Grundzüge forensischer Sprachverarbeitung.* Kriminalistik, Heidelberg, 1987.

H.J. Künzel. Field procedures in forensic speaker recognition. In *Studies in General and English Phonetics - Essays in Honour of J.D. O'Connor.* Routledge, London, 1995.

H.J. Künzel. Effects of voice disguise on speaking fundamental frequency. *Forensic Linguistics: The International Journal of Speech, Language and the Law*, 7(2):149–179, 2000.

H.J. Künzel, J. González-Rodríguez, and J. Ortega-García. Effect of voice disguise on the performance of a forensic automatic speaker recognition system. In *Proceedings of the ODYSSEY - The Speaker and Language Recognition Workshop*, pages 153–156, Toledo, Spain, 2004.

P. Ladefoged. Linguistic aspects of respiratory phenomena. *Annals of the New York Academy of Sciences*, 155:141–151, 1968.

P. Ladefoged. *Vowels and Consonants - An introduction to the Sounds of Languages.* Blackwell, Oxford, 2001.

N.J. Lass, D.S. Trapp, M.K. Baldwin, K.A. Scherbick, and D.L. Wright. Effect of vocal disguise on judgments of speakers' sex and race. *Perceptual and Motor Skills*, 54(3 Pt 2):1235–40, June 1982.

Y.W. Lau, D. Tran, and M. Wagner. Testing voice mimicry with the YOHO speaker verification corpus. In *Knowledge-Based Intelligent Information and Engineering Systems*, volume 3684 of *Lecture Notes in Computer Science*, pages 15–20, Springer, Heidelberg, 2005.

Y.W. Lau, M. Wagner, and D. Tran. Vulnerability of speaker verification to voice mimicking. In *Proceedings of the International Symposium on Intelligent Multimedia, Video and Speech Processing*, Hong Kong, 2004.

J. Laver. *Principles of Phonetics.* Cambridge University Press, Cambridge, 1994.

D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems: Proceedings of the 2000 Conference*, volume 13, pages 556–562. MIT Press, 2001.

S. Lee, A. Potamianos, and S. Narayanan. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *Journal of the Acoustical Society of America*, 105(3), March 1999.

I. Lehiste. *Suprasegmentals.* MIT Press, Cambridge, MA, 1970.

X. Li, J. Tao, M.T. Johnson, J. Soltis, A. Savage, K.M. Leong, and J.D. Newman. Stress and emotion classification using jitter and shimmer features. In *Proceedings of the ICASSP*, volume 4, pages 1081–1084, Honolulu, Hawaii, April 2005.

J. Lindberg and M. Blomberg. Vulnerability in speaker verification. A study of technical impostor techniques. In *Proceedings of the Eurospeech*, pages 1211–1214, Budapest, Hungary, 1999.

Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, 1980.

S.E. Linville. *Vocal Aging.* Singular, San Diego, CA, 2001.

S. Lucey and T. Chen. Improved audio-visual speaker recognition via the use of a hybrid combination strategy. In *Proceedings of the Fourth International Conference on Audio- and Video- Based Biometric Person Authentication*, Guildford, UK, 2003.

R.C. Lummis. Speaker verification by computer using speech intensity for temporal registration. *IEEE Transactions on Audio and Electroacoustics*, 21(2):80–89, 1973.

J. Lüttin and G. Maître. Evaluation protocol for the Extended M2VTS database (XM2VTSDB). IDIAP Communication 05, IDIAP, Martigny, Switzerland, 1998.

E. Machlin. *Dialects for the stage.* Routledge/Theater Arts, New York, 1975.

D. Maltoni, D. Maio, A.K. Jain, and S. Prabhakar. *Handbook of Fingerprint Recognition.* Springer, New York, 2003.

J.D. Markel and S.B. Davis. Text-independent speaker identification from a large linguistically unconstrained time-spaced data base. In *Proceedings of the ICASSP*, 1978.

J.D. Markel, B.T. Oshika, and A.H. Gray Jr. Long-term feature averaging for speaker recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25(4):330–337, 1977.

D. Markham. *Phonetic Imitation, Accent, and the Learner.* PhD thesis, Lund University, Lund, 1997.

A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Eurospeech*, volume 4, pages 1895–1898, Rhodos, Greece, 1997.

M. Mashimo, T. Toda, H. Kawanami, H. Kashioka, K. Shikano, and N. Campbell. Evaluation of cross-language voice conversion using bilingual and non-bilingual databases. In *Proceedings of the ICSLP*, 2002.

M. Mashimo, T. Toda, K. Shikano, and N. Campbell. Evaluation of cross-language voice conversion based on GMM and STRAIGHT. In *Proceedings of the Eurospeech*, Aalborg, Denmark, 2001.

T. Masuko, K. Tokuda, and T. Tobayashi. Imposture using synthetic speech against speaker verification based on spectrum and pitch. In *Proceedings of the ICSLP*, Beijing, China, 2000.

D. Matrouf, J.F. Bonastre, and C. Fredouille. Effect of speech transformation on impostor acceptance. In *Proceedings of the ICASSP*, Tolouse, France, 2006.

T. Matsui and S. Furui. Likelihood normalization for speaker verification using a phoneme- and speaker-independent model. *Speech Communication*, 17:109–116, 1995.

R.J. McAulay and T.F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4), 1986.

D. Michaelis, M. Fröhlich, H.W. Strube, E. Kruse, B. Story, and I.R. Titze. Some simulations concerning jitter and shimmer measurement. In *3rd International Workshop on Advances in Quantitative Laryngoscopy*, Aachen, Germany, 744-754 1998.

N. Minematsu, M. Sekiguchi, and K. Hirose. A perceptual study of speaker age. In *Proceedings of the ICASSP*, pages 123–140, 2002.

S. Moosmuller. Phonological variation in speaker identification. *Forensic Linguistics*, 3(1):29–47, 1997.

C. Nadeu, J. Hernando, and M. Gorricho. On the decorrelation of filter bank energies in speech recognition. In *Proceedings of the Eurospeech*, pages 1381–1384, 1995.

J. Navrátil, Q. Jin, W. Andrews, and J. Campbell. Phonetic speaker recognition using maximum likelihood binary decision tree models. In *Proceedings of the ICASSP*, 2003.

R.G. Newcombe. Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine*, 17(8):857–872, April 1998.

S. Nooteboom. *The Prosody of Speech: Melody and Rhythm. The Handbook of Phonetic Sciences*, pages 641–673. Blackwell Publishers Ltd, Oxford, 1997.

J. Oglesby and J.S. Mason. Optimization of neural models for speaker identification. In *Proceedings of the ICASSP*, pages 261–264, 1990.

A.V. Oppenheim. From frequency to quefrency: A history of the cepstrum. *IEEE Signal Processing Magazine*, September 2004.

J. Pelecanos and S. Sridharan. Feature warping for robust speaker verification. In *Proceedings of the Odyssey Speaker Recognition Workshop*, pages 213–218, Crete, Greece, 2001.

B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D.A. Reynolds, and B. Xiang. Using prosodic and conversational features for high-performance speaker recognition: Report from JHU WS'02. In *Proceedings of the ICASSP*, pages 792–795, Hong-Kong, April 2003.

J.W. Picone. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9): 1215–1247, 1993.

J. Pittam. *Voice in Social Interaction; an Interdisciplinary Approach.* SAGE Publications, Thousand Oaks, 1994.

L.A. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

L.R. Rabiner and B.H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3:4–16, January 1986.

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

L.R. Rabiner and W. Schafer. *Digital Processing of Speech Signal.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1978.

A.R. Reich. Detecting the presence of vocal disguise in the male voice. *Journal of the Acoustical Society of America*, 69(5):1458–1460, May 1981.

D.A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17:91–108, 1995.

D.A. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *Proceedings of the Eurospeech*, Rhodes, Greece, September 1997.

D.A. Reynolds, W. Andrews, J. Campbell, J. Navrátil, B. Peskin, A. Adami, Q. Jin, D. Klusáček, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang. Exploiting high-level information for high-performance speaker recognition (SuperSID Project Final Report). Technical report, MIT Lincoln Laboratory, US Department of Defense, IBM, International Computer Science Institute, Oregon Graduate Institute, Carnegie Mellon University, Charles University, York University, Princeton University, Cornell University, 2002.

D.A. Reynolds, W. Andrews, J. Campbell, J. Navrátil, B. Peskin, A. Adami, Qin Jin, D. Klusáček, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and Bing Xiang. The SuperSID project: exploiting high-level information for high-accuracy speaker recognition. In *Proceedings of the ICASSP*, 2003.

D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10:19–41, 2000.

D.A. Reynolds and R.C. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3:72–83, 1995.

R.D. Rodman. Speaker recognition of disguised voices. In M. Demirekler, A. Saranli, H. Altincay, and A. Paoloni, editors, *Proceedings of the Consortium on Speech Technology Conference on Speaker Recognition by Man and Machine: Directions for Forensic Application*, pages 9–22, Ankara, Turkey, 1998. COST250 Publishing Arm.

P. Rose. *Forensic Speaker Identification.* Taylor & Francis Ltd, 2002.

A.E. Rosenberg and S. Parthasarathy. Speaker background models for connected digit password speaker verification. In *Proceedings of the ICASSP*, 1996.

A. Ross and R. Govindarajan. Feature level fusion using hand and face biometrics. In *Proceedings of the SPIE Conference on Biometric Technology for Human Identification*, Orlando, FL, March-April 2005.

L. Rudasi and S.A. Zahorian. Text-independent talker identification with neural networks. In *Proceedings of the ICASSP*, pages 389–392, 1991.

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49, 1978.

N. Scheffer, J.F. Bonastre, A. Ghio, and B. Teston. Gémellité et reconnaissance automatique du locuteur. In *Proceedings of the XXVth Journées d'Etude sur la Parole*, Fes, Morocco, 2004.

A. Schmidt-Nielsen and T.H. Crystal. Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 Speaker Evaluation Data. *Digital Signal Processing*, 10:249–266, 2000.

S. Schötz. A perceptual study of speaker age. In *Proceedings of the Fonetik 2001*, Lund, 2001. Lund Working Papers.

E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke. Modeling prosodic feature sequences for speaker recognition. *Speech Communication*, 46(3-4):455–472, 2005.

E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32:127–154, 2000.

M. Shridhar, N. Mohankrishnan, and M. Baraniecki. Text-independent speaker recognition using orthogonal linear prediction. In *Proceedings of the ICASSP*, pages 333–336, Detroit, MI, May 1981.

R.W. Shuy. Dialect as evidence in law cases. *Journal of English Linguistics*, 23(1):195–208, 1995.

M. Skosan and D. Mashao. Modified segmental histogram equalization for robust speaker verification. *Pattern Recognition Letters*, 27(5):479–486, April 2006.

R.E. Slyh, W.T. Nelson, and E.G. Hansen. Analysis of mrate, shimmer, jitter, and F0 contour features across stress and speaking style in the SUSAS database. In *Proceedings of the ICASSP*, pages 2091–2094, 1999.

R. Snelick, M. Indovina, J. Yen, and A. Mink. Multimodal biometrics: issues in design and testing. In *Proceedings of the Fifth International Conference on Multimodal Interfaces*, pages 68–72, Vancouver, Canada, 2003.

R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain. Large-scale evaluation of multimodal biometric authentication using state-of-the art systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):450–455, March 2005.

K. Sönmez, L. Heck, M. Weintraub, and E. Shriberg. A lognormal tied mixture model of pitch for prosody-based speaker recognition. In *Proceedings of the Eurospeech*, pages 1391–1394, Greece, 1997.

K. Sönmez, E. Shriberg, L. Heck, and M. Weintraub. Modeling dynamic prosodic variation for speaker verification. In *Proceedings of the ICSLP*, 1998.

F.K. Soong, A.E. Rosenberg, L.R. Rabiner, and B.H. Juang. A vector quantization approach to speaker recognition. In *Proceedings of the ICASSP*, pages 387–390, 1985.

K.N. Stevens, C.E. Williams, J.R. Carbonelli, and B. Woods. Speaker authentication and identification: a comparison of spectrographic and auditory presentations of speech material. *Journal of the Acoustical Society of America*, 43:1596–1607, 1968.

S.S. Stevens. The mel scale equates the magnitude of perceived differences in pitch at different frequencies. *Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

Y. Stylianou. *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification.* PhD thesis, École Nationale Supérieure des Télécommunications, Paris, France, January 1996.

Y. Stylianou, O. Cappé, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, 1998.

K.P.H. Sullivan and J. Pelecanos. Revisiting Carl Bildt's impostor: Would a speaker verification system foil him? In *Proceedings of the 3rd International Conference on Audio- and Video-Based Biometric Person Authentication*, volume 2091, pages 144–149, Halmstad, Sweden, 2001.

D. Sündermann, H. Höge, A. Bonafonte, H. Ney, and J. Hirschberg. TC-Star: Cross Language Voice Conversion Revisited. In *Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, 2006.

O. Tosi, H. Over, W. Lashbrook, C. Pedrev, J. Nicol, and E. Nash. Experiment on voice identification. *Journal of the Acoustical Society of America*, 51:2030–2043, 1972.

J. Tusón(dir.). *Diccionari de lingüística.* Vox, Barcelona, 2000.

I. Wagner. A new jitter-algorithm to quantify hoarseness: an exploratory study. *Forensic Linguistics*, 2:18–27, 1995.

Y. Wang, Y. Wang, and T. Tan. Combining fingerprint and voiceprint biometrics for identity verification: an experimental comparison. In *Proceedings of the International Conference on Biometric Authentication*, volume 3072 of *Lecture Notes in Computer Science*, pages 663–670, Hong Kong, China, 2004.

F. Weber, L. Manganaro, B. Peskin, and E. Shriberg. Using prosodic and lexical information for speaker identification. In *Proceedings of the ICASSP*, volume 1, pages 141–144, 2002.

A. Wennerstrom. *The Music of Everyday Speech. Prosody and Discourse Analysis.* Oxford University Press, 2001.

S. Werner and E. Keller. Prosodic aspects of speech. In E. Keller, editor, *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges*, pages 23–40. John Wiley, Chichester, New York, 1994.

H.F. Wertzner, S. Schreiber, and L. Amaro. Analysis of the fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders. *Brazilian Journal of Otorhinolaryngology*, 71(5):582–588, September-October 2005.

C.E. Williams and K.N. Stevens. Emotions and speech: some acoustical correlates. *Journal of the Acoustical Society of America*, 52:1238–1250, 1972.

F. Wittig and C. Müller. Implicit feedback for user-adaptive systems by analyzing the user's speech. In *Proceedings of the ABIS-03 (Adaptivität und Benutzermodellierung in interaktiven Softwaresystemen)*, Karlsruhe, Germany, 2005.

J.J. Wolf. Efficient acoustic parameters for speaker recognition. *Journal of the Acoustical Society of America*, 51:2044–2056, 1972.

E.H. Wrench. A real time implementation of a text independent speaker recognition system. In *Proceedings of the ICASSP*, 1981.

H. Ye and S. Young. Quality-enhanced voice morphing using maximum likelihood transformations. *IEEE Transactions on Audio, Speech and Language Processing*, 2006.

S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas. Exploiting discriminant information in non-negative matrix factorization with application to face verification. *IEEE Transactions on Neural Networks*, 2005a.

S. Zafeiriou, A. Tefas, and I. Pitas. Discriminant NMF-faces for frontal face verification. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, Mystic, Connecticut, 2005b.

E. Zetterholm. The role of prosody in voice imitation. In *Proceedings of the VIII Conference on Nordic Prosody*, pages 239–252, 2000.

E. Zetterholm. A comparative survey of phonetic features of two impersonators. In *Proceedings of Fonetik*, volume 44, pages 129–132, 2002a.

E. Zetterholm. Intonation pattern and duration differences in imitated speech. In *Proceedings of Speech Prosody*, 2002b.

E. Zetterholm. *Voice Imitation. A phonetic study of perceptual illusions and acoustic success.* PhD thesis, Department of Linguistics and Phonetics, Lund University, Lund, 2003.

E. Zetterholm. Same speaker - different voices. a study of one impersonator and some of his different imitations. In *Proceedings of the 11th Australian International Conference on Speech Science and Technology*, pages 70–75, Auckland, New Zealand, December 2006.

E. Zetterholm, M. Blomberg, and D. Elenius. A comparison between human perception and a speaker verification system score of a voice imitation. In *Proceedings of the 10th Australian International Conference on Speech Science and Technology*, pages 393–397, Sydney, Australia, December 2004.

E. Zetterholm, K.P.H. Sullivan, and J. van Doorn. The impact of semantic expectation on the acceptance of a voice imitation. In *Proceedings of the 9th Australian International Conference on Speech Science and Technology*, 2002.

W. Zhao, R. Chellapa, A. Rosenfeld, and P.J. Phillips. Face recognition: A literature survey. *ACM Computing Surveys*, pages 399–458, 2003.

List of author's publications

- M. Farrús, J. Anguita, X. Anguera, J.M. Crego, A. de Gispert, J. Hernando, and C. Nadeu. Els sistemes de reconeixement de veu i traducció automàtica en català: present i futur. In *Actes del II Congrés d'Enginyeria en Llengua Catalana*, Andorra la Vella, Andorra, 2004.

- X. Anguera, Farrús, and J. Hernando. Segmentació de locutor per a la indexació automàtica de bases de dades multimèdia en català. In A*ctes del II Congrés d'Enginyeria en Llengua Catalana*, Andorra la Vella, Andorra, 2004.

- R. Cerdà, M. Farrús, and J. Hernando. Hacia una sinergia metodológica en la identificación de locutores. In *Filología y Lingüística. Estudios ofrecidos a Antonio Quilis.*, volume 2, pages 1515-1528, CSIC, Madrid, 2005.

- M. Farrús, J. Anguita, J. Hernando, and R. Cerdà. Fusión de sistemas de reconocimiento basados en características de alto y bajo nivel. In *Actas del III Congreso de la Sociedad Española de Acústica Forense*, Santiago de Compostela, October 2005.

- R. Cerdà, M, Farrús, J. Hernando, and M. Veyrat. Un experimento sobre la noción de campo de dispersión fonológica. In *Actas del III Congreso de la Sociedad Española de Acústica Forense*, Santiago de Compostela, 2005.

- J. Luque, R. Morros, A. Garde, A. Anguita, M. Farrús, D. Macho, F. Marqués, C. Martínez, V. Vilaplana, and J. Hernando. Audio, video and multimodal person identification in a smart room. In *Proceedings of the CLEAR Workshop*, Southampton, volume 4122 of *Lecture Notes in Computer Science*, pages 258-269, 2006.

- J. Hernando, M. Farrús, A. Garde, and P. Ejarque. Person verification by fusion of prosodic, voice spectral and facial parameters. In *Proceedings of the International Conference on Security and Cryptography*, pages 17-23, Setúbal, Portugal.

- M. Farrús, E. Eriksson, K.P.H. Sullivan, and J. Hernando. Dialect imitations in speaker recognition. In Proceedings of the 2n European IAFL Conference on Forensic Linguistics, Language and the Law, pages 347-353, Barcelona, 2006.

- M. Farrús, A. Garde, P. Ejarque, J. Luque, and J. Hernando. On the fusion of prosody, voice spectrum and face features for multimodal person verification. In *Proceedings of the ICSLP*, pages 2106-2109, Pittsburgh, 2006.

- J. Luque, R. Morros, J. Anguita, M. Farrús, D. Macho, F. Marqués, C. Martínez, V. Vilaplana, and J. Hernando. Multimodal person identification in a smart room. In *Actas de las IV Jornadas en Tecnología del Habla*, pages 327-331, Zaragoza, 2006.

- M. Farrús, E. Eriksson, K.P.H. Sullivan, and J. Hernando. Speaker recognition and accents. In *Proceedings of the Femte Svenska Lingvistikkonferensen*, Umeå, 2007.

- M. Farrús, J. Hernando, and P. Ejarque. Jitter and shimmer measurements for speaker recognition. In *Proceedings of the Eurospeech*, pages 778-781, Antwerp, Belgium, 2007.

- M. Farrús, P. Ejarque, A. Temko, and J. Hernando. Histogram equalization in SVM multimodal person verification. In *Proceedings of the IEEE International Conference on Biometrics*, Seoul, Korea, volume 4642 of *Lecture Notes in Computer Science*, pages 819-827, August 2007.

- M. Farrús, M. Wagner, J. Anguita, and J. Hernando. How vulnerable are prosodic features to professional imitators? In *Proceedings of Odyssey - The Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, 2008.

- M. Farrús, M. Wagner, J. Anguita, and J. Hernando. Robustness of prosodic features to voice imitation. In *Proceedings of the Interspeech*, Brisbane, Australia, 2008.

- M. Farrús, D. Erro, and J. Hernando. Speaker recognition robustness to voice conversion. In *Proceedings of the IV Jornadas de Reconocimiento Biométrico de Personas*, Valladolid, Spain, 2008.