TÉCNICO LISBOA

Anti-spoofing: Speaker verification vs. voice conversion

Maria Joana Ribeiro Folgado Correia

Thesis to obtain the Master of Science Degree in

Engineering Physics

Supervisors: Prof. Dr. Isabel Maria Martins Trancoso

Prof. Dr. Horácio João Matos Fernandes

Examination Committee

Chairperson: Prof. Dr. Maria Teresa Haderer de la Peña Stadler

Supervisor: Prof. Dr. Isabel Maria Martins Trancoso

Member of the Committee: Prof. Dr. Ana Luísa Nobre Fred

April 2014

To my mother.

Acknowledgments

My deepest gratitude goes to Prof. Isabel Trancoso who has given me tremendous support during this thesis. Her support has greatly exceeded the technical insights and suggestions to improve the quality of this thesis. The countless times she stood up for me will make it impossible for me to ever thank her enough.

Most of the technical quality of this work I owe to Prof. Alberto Abad, who was also my supervisor. Bureaucratic issues prevent his official acknowledgment as such, despise his numerous contributions to this work. He was always available to spend time with me and my many questions, making useful and insightful suggestions about the approaches that were being taken or the experiments being performed. His guidance was invaluable and without him, I am sure that this thesis would not be half as good. So thank you!

I would also like to acknowledge Prof. Horácio Fernandes for allowing this thesis to happen.

I would like to thank Zhizheng Wu and Prof. Haizhou Li of the Nanyang Technological University, Singapore for providing the GMM-based and unit selection-based converted speech corpora, and for their helpful suggestions.

My colleagues, the researchers and the Professors of L^2F in general, deserve being mentioned here, for collectively making of this laboratory such a nice place to work and for nourishing a friendly environment.

On a personal note, I must thank my mother, my brother and my grandparents for all the education, support and love which I will never be able to properly repay. To Mónica, Carlos and Luis, I thank them their everlasting presence in my life. Finally, to João, I thank him for everything.

Resumo

As técnicas de conversão de voz, que modificam a voz de um orador para que se assemelhe a um outro, apresentam uma ameaça para sistemas automáticos de verificação do orador. Nesta tese, é avaliada a vulnerabilidade destes sistemas a ataques de *spoofing* realizados por voz convertida. De forma a superar estes ataques, são implementados detectores de fala convertida baseados em parâmetros de curta e longa duração. Adicionalmente, são propostos novos detectores de fala convertida que utilizam uma representação compacta de parâmetros, a par com técnicas de modelação discriminativas. São realizadas experiências que visam a fusão de pares de detectores de fala convertida usando parâmetros de curta e longa duração com o objectivo de melhorar o desempenho conjunto dos detectores. Os resultados indicam que a fusão dos detectores de fala convertida propostos superam os existentes, atingindo uma taxa de detecção de 98,1% para fala natural e de 97,6% para fala convertida. Os detectores de fala convertida propostos são incluídos como mecanismo de *anti-spoofing* no sistema de verificação do orador, sendo aplicados como módulo de pós-processamento a todas as locuções aceites como pertencentes a um orador alvo pelo sistema. O desempenho do sistema de verificação do orador com o mecanismo de *anti-spoofing* é reavaliado e comparado ao de um mecanismo ideal. Os resultados mostram que o sistema de verificação do orador volta a ter um desempenho considerado normal, em que apenas 1,9% dos segmentos de fala rejeitados continuam mal classificados.

Termos chave

Verificação do orador, conversão de voz, *anti-spoofing*, biometria, processamento da fala

Abstract

Voice conversion (VC) techniques, which modify a speaker's voice to sound like another's, present a threat to automatic speaker verification (SV) systems. In this thesis, the vulnerability of a state-of-the-art SV system against converted speech spoofing attacks is evaluated. To overcome the spoofing attacks, state-of-the-art converted speech detectors based on short- and long-term features are implemented. Additionally, new converted speech detector using a compact feature representation and a discriminative modeling approach are proposed. Experiments on pairing converted speech detectors based on short- and long-term features are made in order to improve the overall performance of the detection task. The results indicate that the proposed fused converted speech detector outperform the state-of-the-art ones, achieving a detection an accuracy of 98.1% for natural utterances and 97.6% for converted utterances. This proposed converted speech detector is used as an anti-spoofing mechanism by the SV system. The mechanism consists of a post-processing module for accepted trials. The performance of the SV system with the anti-spoofing mechanism is reevaluated and compared to that of an ideal mechanism. The results show that the SV system's performance returns to acceptable values, with an average of 1.9% of the rejected trails remaining misclassified.

Keywords

Speaker verification, voice conversion, anti-spoofing, biometrics, speech processing

Index

List of tables

List of figures

List of acronyms

FAR	False Acceptance Rate
GDCC	Group Delay Cepstral Coefficient
GDF	Group Delay Function
GMM	Gaussian Mixture Model
MFCC	Mel-Frequency Cepstral Coefficient
MGDCC	Modified Group Delay Cepstral Coefficient
MGDF	Modified Group Delay Function
MM	Magnitude Modulation
PM	Phase Modulation
SV	Speaker Verification
SVM	Support Vector Machine
VAD	Voice Activity Detection
VC	Voice Conversion

1 Introduction

1.1 Motivation

Speech is an inherently important human communication tool. As such, the development of computational systems that process speech in various ways is a very interesting and important challenge. Such systems have received widespread attention over the last decades and have a broad range of applications. Some of those include systems that concentrate on the intelligible content of speech, such as speech recognition; others focus on the timbral quality of speech, for instance speaker identification; and even others that refer to signal synthesis such as voice synthesis. This thesis focuses on two applications in particular: Speaker verification (SV) and voice conversion (VC).

SV refers to the task of verifying the identity of an unknown person as being a previously defined person solely from their voice [1]. This is possible because no two individuals sound identical. Their vocal tract shapes, larynx sizes and other parts of their voice production organs are different. In addition to these physical differences, each speaker has his or her characteristic manner of speaking, including the use of a particular accent, rhythm, intonation style, and pronunciation pattern.

SV systems explore these differences to automatically classify a speaker. These systems exist since the 1960s, when the first system of this kind was created. It was based on the analysis of X-rays on individuals making specific sounds [2]. Since then, SV systems have significantly evolved, and state-of-the-art techniques often include machine learning approaches for speaker modeling and decision. It is not uncommon to find SV systems with error rates as low as 1%, sometimes even less.

VC is a much more recent technique. It aims at modifying the one speaker's voice characteristics, the *source* speaker, into sounding as if they belonged to another speaker, the *target* speaker, without changing the linguistic contents of the converted utterance [3]. It takes into account both the *timbre* and the *prosody* of the source and target speakers.

For many years, successful VC has been difficult to achieve, particularly in terms of the perceived quality and naturalness of the converted speech. However, recent progress has enabled the integration of VC into a number of applications. In parallel, this progress has also enabled the use of VC to transform the voice of a source user into sounding like a target user, who himself is the target speaker of an SV system. Thus the converted speech can be used to try to fool an SV system and perform what is known as a *spoofing attack*.

The vulnerability of SV systems against spoofing attacks has been widely recognized [4][5]. Given the quality of converted speech using state-of-the-art techniques, an attack of such nature often results in a dramatic increase of the false acceptance rate (FAR) of the SV system performance, which creates an important security breach in SV systems.

The first efforts in order to make an SV system more robust against converted speech spoofing attacks are very recent. The most successful approaches are based on the use of converted speech detection modules that discriminate between natural and synthetic speech [6-10].

The possibility of further developing anti-spoofing mechanisms for SV systems has served as the main motivation for this thesis.

1.2 Contributions

The main contribution of this thesis was the development of an anti-spoofing mechanism based on a new converted speech detector. All the converted speech detectors reported in the literature use Gaussian mixture models (GMMs) as the modeling technique. Given that converted speech detection is a binary task, this thesis proposes the use of a discriminative model for the detection task, based on a support vector machine (SVM) model. Additionally, the proposed detectors use a new way of representing the features, which is much lighter and makes the trained models less complex. The combination of the proposed modeling technique and the new feature representation for converted speech detection allowed significant improvements in the performance over the state-of-the-art.

1.3 Document structure

This thesis is divided into seven chapters. In Chapter 2, the historical and theoretical fundaments of the relevant speech processing tasks are briefly reviewed. The main focus is on speech feature extraction, SV and VC.

Chapter 3 describes the experiments related to the implementation of the state-of-the-art SV system. These experiments establish a baseline performance and confirm the vulnerability of SV systems against converted speech spoofing attacks.

This vulnerability provides the motivation for the review of the literature on the topic of anti-spoofing mechanisms which is the focus of Chapter 4.

Chapter 5 starts by describing the experiments related to the implementation of existing converted speech detectors. The emphasis of this chapter is on the proposal of new converted speech detectors, including a comparison with the previously implemented detectors. The chapter also features experiments on the fusion of the converted speech detectors by pairing detectors using short- and long-term features.

In Chapter 6, the proposed converted speech detector is integrated as an anti-spoofing mechanism for SV systems. The performance of the SV system is reevaluated in this chapter.

Finally, Chapter 7 draws the main conclusions and provides some suggestions for future work.

2 Speaker verification and voice conversion

2.1 Introduction

The main goal of this thesis is to develop more sophisticated anti-spoofing mechanisms that can mitigate the performance issues that converted speech spoofing attack pose to SV systems. In order to so it is important to be aware of the theoretical fundaments of the speech processing related topics. As such, this chapter focuses on making a review of the literature in terms of:

- Features – the notion of feature; the characteristics to take into account when choosing them; the *de facto* standard feature: the Mel-frequency cepstral coefficient;
- SV systems – a brief overview of the typical SV system; the de facto standard of modeling techniques in SV: the GMM; other modeling techniques recently introduced in SV; decision methods; and evaluation metrics;
- VC systems – a brief overview of the concept of VC; the de facto standards in VC: GMM-based VC; a popular alternative approach to GMM-based voice conversion: unit selection-based voice conversion; and evaluation metrics

2.2 Information representation and feature extraction

In speech processing related tasks, and in particular in SV and VC systems, the speech signal is not usually fed directly to the system. Instead, the speech signal is transformed into a more compact and meaningful representation, suitable to be processed. This representation, consisting in a set of features extracted from the speech signal, is then used by the application in question.

The features should have some characteristics to ensure that they contain important and accurate information about it [11]:

1. Practical, in the sense that they occur naturally and frequently in speech and they should be easy enough to measure;
2. Robust, meaning that the ideal features should be invariant over time (speaker's life) nor should they be affected by the speaker's health, and they should also be, as much as possible, unaffected by environmental noise or channel characteristics;
3. Secure, so they are not subject to mimicry.

In practice the ideal feature with all the mentioned attributes does not exist, but from previous experiments it has been found that features derived from the magnitude spectrum of the speech signal are the most successful [11]. These features contain information on low-level cues that are related to the acoustic aspects of speech as well as glottal and vocal-tract characteristics.

Many different features have been investigated in the literature. Linear prediction coefficients (LPC) [12] have received special attention in this regard as they are directly derived from the

speaker's speech production model and perceptual linear prediction (PLP) coefficients [13], which are also based on human perception and auditory processing are another broadly used feature. However, over the past two decades spectral based features, mostly derived by direct application of the Fourier Transform, have become the most popular. The MFCCs are considered the *de facto* standard of such features. These make use of a Mel-scale which approximates the human auditory system's response. Given its popularity, they were used in this thesis as a baseline feature for the applications implemented. The next section explains how to derive MFCCs from a speech signal.

2.2.1 Mel-frequency cepstral coefficients

The MFCCs are the coefficients that collectively make up the Mel-frequency cepstrum, which in its turn, is a representation of the short-term power spectrum of the speech signal. The MFCCs carry information about the short-term magnitude spectrum of the speech signal.

In order to derive the MFCCs, it is necessary to process the signal using short time Fourier analysis. For this, the signal has to be processed in time windows short enough so that each one can be considered quasi-stationary. Typically this means using 5 to 25ms windows. For a given speech signal window, $x(n)$, its STFT is:

$$X(\omega) = |X(\omega)|e^{j\phi(\omega)}, \tag{1}$$

where $|X(\omega)|$ is the magnitude spectrum and $\phi(\omega)$ is the phase spectrum. The power spectrum is defined as $|X(\omega)|^2$.

Pre-processing: Before deriving the MFCCs it is necessary to pre-process the speech signal and separate it in windows, to allow the short-term Fourier analysis. The typical pre-processing framing process is depicted in Fig. 2.1 and it is as follows:

1. Pre-emphasis the whole speech signal: usually a first order high-pass filter is applied to the waveform to emphasize high frequencies and compensate for the human speech production process which tends to attenuate those frequencies;
2. Framing: the speech signal is divided into overlapped, fixed duration segments called frames. These frames have small durations, up to 25ms, and update rates half or less the window size, resulting in an overlap of at least 50% between consecutive frames;
3. Windowing: each frame is multiplied by a window function (a Hamming window is the most popular option) which is needed to smooth the effect of using a finite-sized segment for the subsequent feature extraction by tapering each frame at the beginning and end edges preventing the spectral artifacts.

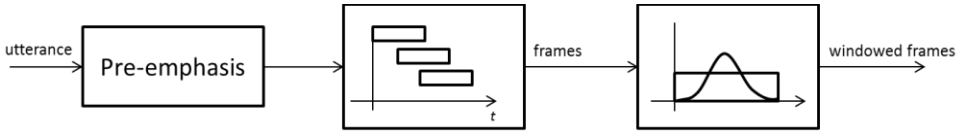

Fig. 2.1 Typical framing processing for a speech signal

Feature extraction: After the pre-processing of the speech signal into windowed frames, the MFCCs are derived for each windowed frame as follows:

1. Fourier Transform: A p-point fast Fourier transform (FFT) operation is applied to the frame. Only the magnitude spectrum is considered and the phase spectrum is discarded;
2. Mel-spaced filter-bank: The p magnitude coefficients are converted to K filter-bank values (acceptable values are, for example, $p = 256$ and $K = 30$, so the spectrum will be smoothed and detailed information will be purposefully lost during the conversion) and a more efficient representation is achieved. Furthermore this can be carried out in a perceptually meaningful way by smoothing logarithmically rather than linearly, specifically using a Mel scale. The filter bank values are derived by multiplying the p magnitude coefficients by the K triangular filter bank weighting functions and then accumulating the results from each filter. The centers of the triangle filter banks are distributed according to a Mel scale, as shown in Fig. 2.2. The conversion of frequencies in a linear scale into a Mel scale can be achieved following:

$$f_{Mel} = 2595 \log_{10}\left(1 + \frac{f_{Lin}}{700}\right); \quad (2)$$

The output of the application of the Mel-spaced filter banks will be K filter bank spectral values, $S_{k\,k=1}^{K}$.

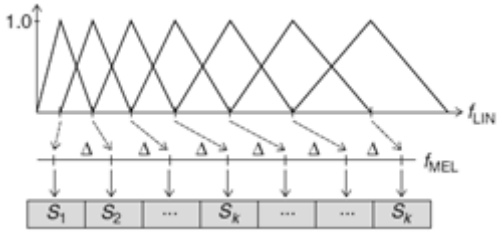

Fig. 2.2 Mel-scale filter bank and filter bank spectral values

3. Cepstral Analysis: To convert the K log filter bank spectral values, $\{log(S_k)\}_{k=1}^{K}$, into L cepstral coefficients, discrete cosine transform (DCT) is used:

$$c_n = \sum_{k=1}^{K} log(S_k) cos\left[n - \left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \quad n = 1, 2, \ldots, L. \quad (3)$$

In contrast with spectral features, which are highly correlated, cepstral features yield a more decorrelated, compact representation. Typically only the first 12 or so MFCCs are kept. Additionally, the special c_0 cepstral coefficient can also be included which, by definition, represents the average log-power of the frame.

The feature extraction process is summarized in Fig. 2.3.

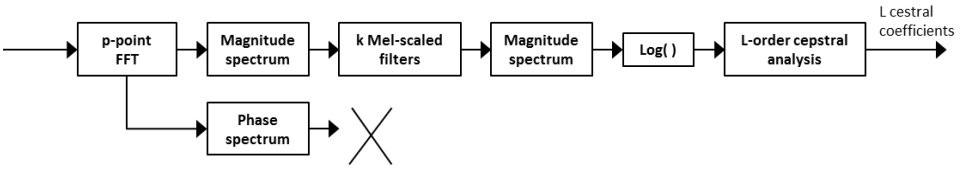

Fig. 2.3 MFCCs feature vector extraction process

4. Concatenation with temporal derivatives: the MFCCs feature vector contains only static information about that frame, which means that there is no dynamic, temporal information between frames, and the coefficients evolution overtime. To overcome that and include some temporal information in the feature vector it is customary to augment the feature vector by including the 1st and 2nd order derivative approximations of the MFCCs. First order derivatives, *delta* or *velocity* parameters are approximated by:

$$\vec{d}_t = \frac{\sum_{p=1}^{P} p\left(\vec{c}_{t+p} - \vec{c}_{t-p}\right)}{2\sum_{p=1}^{P} p^2}, \tag{4}$$

where typically $p = 2$. By replacing $\vec{c}_t$ by $\vec{d}_t$ one can similarly derive the second order derivative, *delta-delta* or *acceleration* parameters. The resulting augmented feature vector has three times the dimension of the original feature vector.

Environment and channel compensation: After deriving the augmented MFCCs feature vectors, it is usual to perform some post-processing to make the features more robust.

Different input devices impose different spectral characteristics on the speech signal (such as bandlimiting and shaping). These characteristics should be removed from the feature vectors in order to allow verification systems to work independently of the input device that is used, so the goal of channel compensation is to remove these channel effects.

A key advantage in transforming spectral magnitude features to log spectral cepstral features is that multiplicative channel and environmental effects become additive [14], this allows a much easier subtraction of these effects through well-known processes like cepstral-mean subtraction (CMN) [15]. The effects of channel compensation can be further augmented by score normalization.

Speech detection: Sometime along the pre-processing, feature extraction or post-processing, it is important to perform some form of voice activity detection (VAD) in order to exclude the frames or the corresponding feature vectors that only contain silence, because they are not relevant to speech related applications. The simplest form of VAD is to set a threshold for the energy of each frame of feature vector and eliminate those that are below the threshold. Moreover, there are other popular and more sophisticated VAD algorithms such as pitch detection [16], or spectral analysis [17], zero crossing rate [18] among others.

2.3 Speaker verification

Biometrics refers to the recognition of humans by their characteristics or traits, and in particular, *speaker recognition* is a form of voice biometrics that aims at recognizing the identity of a given person solely by his voice characteristics. It is important to emphasize that in speaker recognition what is recognized is the identity of the speaker (the *who*) and not the content of what was spoken (the *what*), that would be *speech recognition*.

Speaker recognition systems have the same kind of applications as other biometric systems, which include recognition technologies in forensics, telephone-based services which require user verification (for example phone banking), information indexation applications that use any form of speech (for example automatically identifying and verifying the identity of the participants in a voice recorded meeting) [19]. For many of these applications, security is a major concern, so the robustness of speaker recognition systems is crucial. This concern has served as a great motivator for the latest evolutions in the field of SV.

2.3.1 Verification vs. identification

There are two major applications of *Speaker Recognition* technologies. If a speaker claims to have a certain identity and the voice is used to verify this claim, it is called *verification* or *authentication*. On the other hand, if the task consists of determining an unknown speaker's identity, it is called *identification*.

For SV the type of match that is performed is a 1:1 match, which means that one speaker claims he is a specific target speaker. To verify this claim, the speech of the unknown person is compared against both the claimed identity and against all other speakers' models (called respectively the target model and the impostor model). The ratio of the two measures is then taken and compared to a threshold. If this ratio is above the threshold the claim is accepted as true, otherwise the claim is rejected. The SV process is summarized in Fig. 2.4.

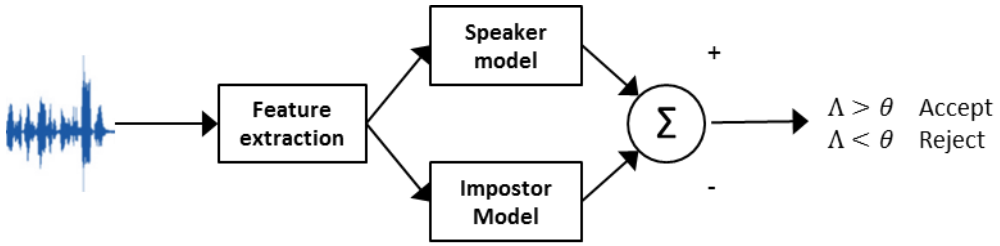

Fig. 2.4 Basic structure of SV system

In the case of speaker identification applications the match is a 1:N where the speaker's voice is compared to a pool of N speaker's models on a given database, as shown in Fig. 2.5. For closed-set identification the speaker is required to be matched with one of the models of the database, contrariwise, if the identification system is of open-set the speaker may be identified as one of the models or as an unknown (a speaker whose model is not contemplated in the database). In this case the first thing to do is to determine whether the speaker belongs to the pool of known speakers, if not, that speaker is considered unknown, otherwise, closed-set identification is carried out.

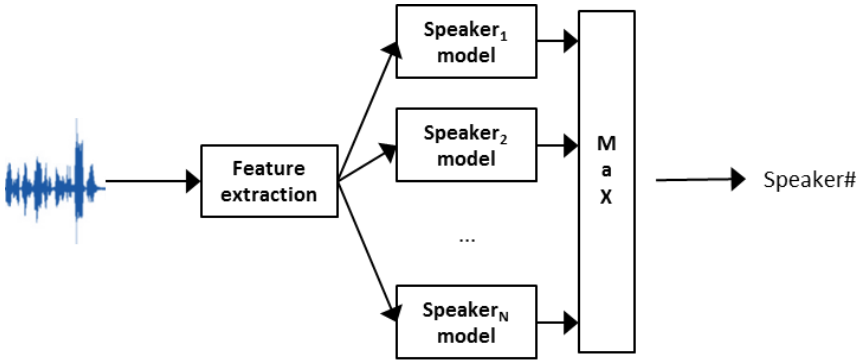

Fig. 2.5 Basic structure of a speaker identification system

2.3.2 Text dependence

A speaker recognition system may or may not depend on *what* the speaker says besides the speaker's identity; such systems are called, respectively, as text-dependent and text-independent speaker recognition systems.

In text-dependent speaker recognition systems the goal is to identify who spoke and also determine what has been said. In a pure text-dependent application the recognition system has prior knowledge of this text (which means a speaker speaks the same text during enrollment and verification phases). Without being as rigid, a text constrained application allows a speaker to use text from a limited vocabulary, such as the digits. The system has prior knowledge of the constrained vocabulary to be used and may have exact knowledge of the text to be spoken, as when using prompted phrases.

Because text-dependent speaker recognition systems require both the words being spoken to be correctly identified and also identifying the correct identity of the speaker they may show improved performances when compared to text-independent speaker recognition systems. The major handicap of text-dependent speaker recognition systems comes from their higher complexity which limits their practical applications. Consequently there is a greater interest for text-independent speaker recognition, which recognizes an individual without any constraints on what the individual is saying (although it's usually assumed that the individual is actually speaking and in a language of interest). In these systems the recognition tends to be more difficult but also more flexible, allowing, for instance, verification of a speaker while he/she is conducting other speech interactions (background verification).

2.3.3 Speaker verification systems

SV systems date back to 1960, when the first system to validate a person's identity based on their voice was built. The model was based on the analysis of X-rays on individuals making specific sounds [2]. With the advancements in technology over next decades, more robust and highly accurate systems have been developed. Presently, state-of-the-art automatic SV systems have a well-defined framework, which comprises two phases: the training phase and the verification phase; and three main blocks: feature extraction, model training and classification, which will be explained in detail in further sections.

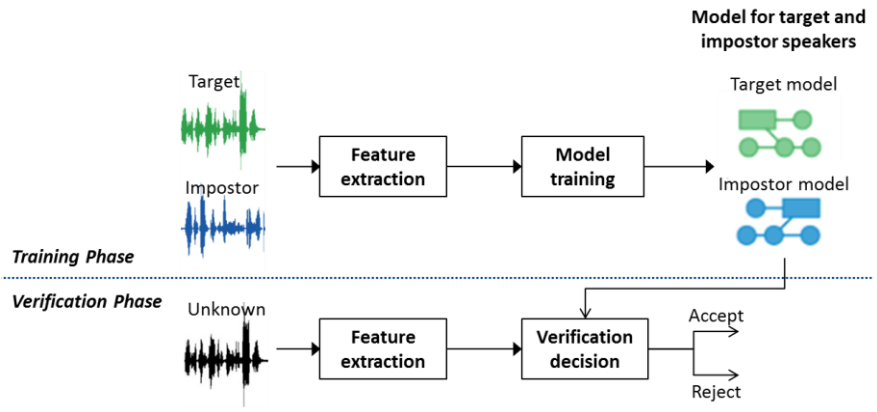

Fig. 2.6 Basic structure of an SV system including training and verification phases

Briefly, an SV system, as Fig. 2.6 shows, works as follows:

In the first phase, the training phase, many speech samples or utterances from different speakers are collected, pre-processed and features characterizing the speech signal are extracted. These features are then used to train models of speakers. For a speaker verification system it is trained a model of the target speaker (using utterances from the target speaker) and, in many cases, a model of the impostor (using utterances from other speakers). Training is done through statistical modeling, where generative, discriminative or mixed approaches are possible. The training phase is followed by a verification phase where the system user provides a test utterance and a claimed identity. The test utterance is processed as those of the enrollment phase and features are extracted (the same type of features as in the previous phase). Afterwards the system makes a decision based on a classifier on whether the test utterance belongs to the target speaker or not.

2.3.3.1 Speaker modeling

The modeling techniques used in SV have suffered fantastic improvements over the last decades. Initially, approaches spanned from human aural and spectrogram comparisons, to simple template matching, to dynamic time-warping approaches. Presently more sophisticated models are estimated resorting to statistical pattern recognition approaches, such as neural networks and Hidden Markov Models (HMMs), GMMs and SVMs.

The SV task consists in taking the decision if an unknown utterance belongs to the user with the claimed identity (the target speaker) or not. Given an utterance X and a claimed identity S the SV system should choose one of the following hypotheses:

H_S: X is pronounced by S

$\overline{H_S}$: X is not pronounced by S

The verification task is accomplished differently depending on the modeling technique that was adopted.

In the following sections brief descriptions of the most prevalent modeling techniques used in SV systems will be given as well as the corresponding techniques to perform the verification of

unknown utterances. Among the described techniques and method, the GMM and GMM-UBM should be highlighted as they are considered the *de facto* standards in SV.

2.3.3.2 Gaussian mixture model

GMMs are the most generic statistical modeling paradigm. This model assumes the feature vectors follow a Gaussian distribution characterized by a mean and a deviation about the mean and that by allowing a mixture of such Gaussians, the distribution of the features of a given speaker may be characterized.

The GMM for speaker j, λ_j, is a weighted sum of M component densities with an output probability for a given feature vector x_t of:

$$p(\boldsymbol{x}_t|\lambda_j) = \sum_{i=1}^{M} \alpha_i \, p_i, \tag{5}$$

where α_i are the mixture weights that satisfy the condition $\sum_{i=1}^{M} \alpha_i = 1$ and p_i are the individual component densities, which for a D-variate Gaussian function are of the form:

$$p_i = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}_t - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_i)}, \tag{6}$$

with the mean vector $\boldsymbol{\mu}_i \in \mathbb{R}^D$(representing the expected spectral feature vector from the state) and the covariance matrix $\boldsymbol{\Sigma}_i \in \mathbb{R}^{D \times D}$ (representing the correlations and variability of spectral features within the state). The GMM for speaker j, λ_j, is parameterized by the mean vectors, covariance matrices and mixture weights from all M component densities:

$$\lambda_j = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i\}_j, \;\; i = 1, 2, \dots, M. \tag{7}$$

As a guideline for the recommended number of mixtures needed for a reliable GMM one can use the number of distinct phones for a given language: for example in the English language there are around 45 phones so a GMM with 64 or more components would be able to represent the individual speaker's broad phonetic class distribution.

The parameter estimation is achieved through an expectation maximization (EM) algorithm [20] which can guarantee monotonic convergence to the set of optimal parameters in few iterations (usually 5 to 10, depending on the problem) [20].

The most common classification method for a GMM-based SV system is maximum likelihood (ML) scoring. For an utterance $X = \{x_1, \dots, x_N\}$, a target model λ_T and an impostor model λ_I the likelihood ratio is:

$$\frac{Pr(X \text{ is from the target speaker})}{Pr(X \text{ is from an impostor})} = \frac{Pr(\lambda_T|X)}{Pr(\lambda_I|X)}. \tag{8}$$

Aplying Bayes' rule and discarding the constant prior probabilites for the target and the impostor, the likelihood ratio in the log domain becomes:

$$\Lambda(X) = \log p(X|\lambda_T) - \log p(X|\lambda_I), \quad (9)$$

Where the term $p(X|\lambda_T)$ is the likelihood that the utterance X belongs to the target speaker and $p(X|\lambda_I)$is the likelihood that the utterance X does not belong to the target speaker. The likelihood ratio is compared with a threshold θ and the target speaker is accepted if $\Lambda(X) \geq \theta$and rejected if $\Lambda(X) < \theta$. The likelihood ratio measures how much better are the target model's scores for the test utterance compared to the impostor model's scores. Then the decision threshold is set to adjust the trade-off between rejecting true target speakers (or miss rate) and accepting false impostor speakers (or FAR). The terms of the likelihood ratio are computed as follows:

$$\log p(X|\lambda_T) = \frac{1}{N}\sum_{n=1}^{N} \log p(x_n|\lambda_T). \quad (10)$$

GMMs are a powerful modeling approch, but suffer two main drawbacks, both related to the quality and quantity of the avaliable training data. One problem comes from the need of enough training data to properly estimate the model parameters. A possible workaround for this is to use diagonal covariance matrices ($\mathbf{\Sigma}_i = diag(\sigma_1^2, \sigma_2^2, \dots, \sigma_D^2)_i \in \mathbb{R}^D$), instead of full matrices which means seting the correlation between MFCCs feature vectors to zero. This is a fair aproximation given that the performance of SV system for this case is not compromised (it may even improve) [21][22]. Aditionaly this "trick" helps reducing computational costs. The second problem with GMMs comes from the fact that this is a generative model. In the generative modeling paradigm the data that is unseen in the training phase and appears in the test data triggers a low score on that data and lowers the overall system performance. The solution for this is simply to be carefull when seleting training data so that the overall train corpus is varied.

2.3.3.3 Gaussian mixture model with universal background model (GMM-UBM)

In the GMM-based SV framework described previously, an unknown test utterance is verified by comparing its score against the target speaker model and against the impostor model. The impostor model is a GMM which models all speakers except the target speaker, and it is commonly referred as the Universal Background Model (UBM). In strict terms the GMM for the impostor model should be trained exclusively with data that did not belong to the target speaker, but in practice this may not happen. In fact, data from all the available speakers may be used, this includes the target speaker, but it must be guaranteed that all the available data is balanced across subgroups (e. g. different gender). The advantage of this approach is that the UBM can be used for SV for any claimed speaker identity. There is also the advantage that UBM, unlike individual speaker GMM, usually don't suffer from problems of insufficient data or unseen data as they use data from many speakers. Since the amount of data used to train a UBM is greatly increased, the GMM parameters are reliably estimated. It is also not uncommon for an UBM to have a larger number of components, for instance, 256 or more. The fundamental concept of the UBM is to represent a speaker-independent distribution of features across all speaker data.

UBMs have other roles in SV besides being used as impostor models. The fact that statistical model's parameters like GMM's parameters can be adapted in the presence of new data using maximum likelihood linear regression (MLLR) or maximum a-posteriori (MAP) adaptation allows the use of a UBM as a starting point to train a speaker model. Thus, an alternative to individual speaker GMM is to train a UBM and then form the speaker GMM by adaptation of the UBM using the individual speaker data as the adaptation data [21].

The adaptation of a UBM for a given speaker is as follows. Given a UBM and training vectors from a speaker, $X = \{x_1, \ldots, x_T\}$, the first thing to do is to determine the probabilistic alignment of the training vectors into the UBM mixture components. For a component i in the UBM and an instant t, that is given by:

$$\Pr(i|x_t) = \frac{w_i\, p_i(x_t)}{\sum_{m=1}^{M} w_m\, p_m(x_t)}, \tag{11}$$

where $p_i(x_t)$ is the density of the mixture component i for the feature vector x_t and w_i is the weight of the mixture component i in the UBM. The probabilistic alignment, $\Pr(i|x_t)$, and the vector x_t are used to compute the sufficient statistics for the weight, mean and variance parameters:

$$n_i = \sum_{t=1}^{T} \Pr\,(i|x_t), \tag{12}$$

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i|x_t)\, x_t, \tag{13}$$

$$E_i(x^2) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i|x_t)\, x_t^2, \tag{14}$$

Finally, the new sufficient statistics of the training data of the speaker are used to update the existing UBM sufficient statistics for a mixture component i, in order to create the adapted parameters for that mixture. The updated follows the equations:

$$\widehat{w}_i = \left[\frac{\alpha_i^w n_i}{T} + (1 - \alpha_i^w) w_i\right] \gamma, \tag{15}$$

$$\hat{\mu}_i = \alpha_i^m E_i(x) + (1 - \alpha_i^m) \mu_i, \tag{16}$$

$$\hat{\sigma}_i^2 = \alpha_i^v E_i(x^2) + (1 - \alpha_i^v)\left(\sigma_i^2 + \mu_i^2\right) - \hat{\mu}_i^2, \tag{17}$$

where $\{\alpha_i^w, \alpha_i^m, \alpha_i^v\}$ are data-dependent adaptation coefficients controlling the balance between old and new estimates for the weights, means and variances, respectively. γ is a scale factor that ensures sum to unity. Many UBM adaptations only consider the adaptation of the mean, setting $\alpha_i^w = 0$ and $\alpha_i^v = 0$. Fig. 2.7 shows an example of the adaptations of UBM parameters to speaker specific data.

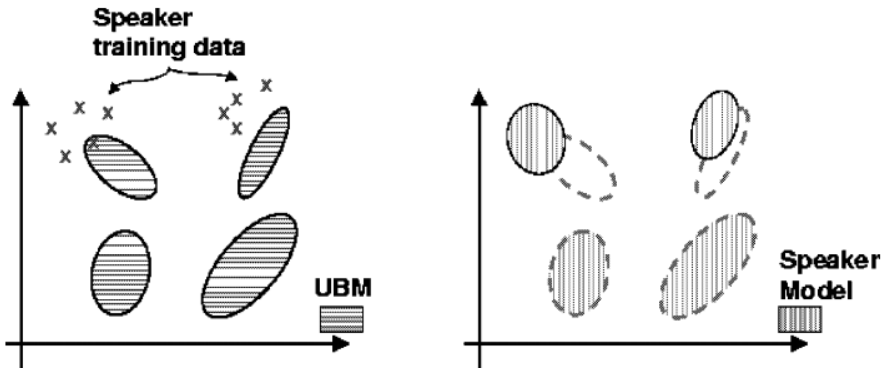

Fig. 2.7 Illustration of the adaptation of UBM parameters with speaker specific data to create a speaker model

The classification of an unknown utterance in a GMM-UBM-based SV system is similar to that of a GMM-based one: a sequence of feature vectors is extracted from the unknown utterance, and those are scored against the target and impostor speaker models. The final decision is made based on ML classifier.

Overall, using GMM-UBM instead of simple GMM for individual speaker modeling has several advantages: Because UBM uses a much larger train corpus than an individually trained GMM the parameters estimated are more reliable even with small amounts of individual speaker adaptation data the resultant GMM-UBM will be equally reliable, much more so when comparing with a directly trained GMM with only the adaptation data [21]. Due to the same reason UBMs require more mixtures than an individually trained GMM, this makes the adapted GMM-UBM (which has the same number of mixtures as the UBM) more capable of handling unseen data: the adaptation algorithm will only modify mixture parameters for which observations exist from the available speaker data and simply copies all others mixture parameters from the UBM. This helps mitigate problems of low scores related to lack of data.

2.3.3.4 GMM-support vector machine (GMM-SVM)

A novel approach in SV, mixing generative with discriminative modeling methods, was proposed by [23] and is based on the generation of a speaker template from a GMM-UBM which can be directly used with an SVM classifier.

Actually, given that SVMs are discriminative classifiers they are a natural solution to SV problems, which are fundamentally two-class problems: either the unknown utterance belongs or does not belong to the target speaker.

In this approach, the starting point is a GMM-UBM, as in Eq. 11, where it is assumed that the covariance matrix is diagonal. For each given speaker utterance, the GMM-UBM training is performed by MAP adaptation of the mixture components means. From the adapted model it is possible to form a GMM supervector by stacking the adapted means of the mixture components into a M-dimensional supervector, where M is the number of mixture components.

The GMM supervector can be seen as a mapping between the utterance and a high dimensional vector. These GMM supervectors can be used as features to be used by an SVM.

The classification of an unknown utterance is given by:

$$f(\boldsymbol{x}) = \sum_{i=1}^{L} \alpha_i t_i K(\boldsymbol{x}, \boldsymbol{v}_i) + k, \tag{18}$$

where t_i are the ideal outputs (for this case it would correspond, for example, $t_i = 1$ for target speaker GMM supervectors and $t_i = -1$ for impostor GMM supervectors), $\boldsymbol{v}_i$ are the support vectors obtained by an optimization process [24] and k is a learned constant. Eq. 18 is subject to the constraints $\sum_{i=1}^{L} \alpha_i t_i = 0$ and $\alpha_i > 0$. The choice of the kernel is subject to Mercer's condition, such that:

$$K(\boldsymbol{x}, \boldsymbol{v}_i) = b(\boldsymbol{x})' b(\boldsymbol{v}_i), \tag{19}$$

where $b(\cdot)$ is the mapping that converts models on the input feature space into supervectors in the expansion space. $f(\boldsymbol{x})$ gives the distance to the separating hyperplane, where the sign of $f(\boldsymbol{x})$ indicates on which side the unknown utterance is. As such, for an unknown utterance's GMM supervectors, $\boldsymbol{x}$, the predicted label is given according to the sign of $f(\boldsymbol{x})$, being 0 if belonging to the negative class or 1 if belonging to the positive class .

2.3.3.5 Factor analysis (FA)

2.3.3.5.1 Joint factor analysis (JFA)

Speaker and session variability between training and test conditions are one of the biggest contributors to the degradation of the performance of GMM-UBM based SV systems [25]. A successful approach to address this problem has been the explicit modeling of the speaker and channel factors through JFA. This technique is based on the decomposition of a speaker dependent GMM supervector, $\boldsymbol{M}$, (where the supervector is defined as the concatenation of the GMM mean vectors), into separate speaker-dependent supervector, $\boldsymbol{s}$, and channel-dependent supervector, $\boldsymbol{c}$:

$$\boldsymbol{M} = \boldsymbol{s} + \boldsymbol{c}, \tag{20}$$

where s and c can be represented by:

$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{V}\,\boldsymbol{y} + \boldsymbol{D}\,\boldsymbol{z}, \tag{21}$$

$$\boldsymbol{C} = \boldsymbol{U}\,\boldsymbol{x}, \tag{22}$$

where, in the speaker dependent component, m is a session speaker independent supervector (extracted from a UBM), $\boldsymbol{V}$ is a low rank matrix representing the primary directions of speaker variability, and $\boldsymbol{D}$ is a diagonal matrix modeling the residual variability not captured by the speaker subspace. The speaker factors $\boldsymbol{y}$, and speaker residuals $\boldsymbol{z}$, are both independent random vectors having standard normal distributions. The channel dependent component is decomposed on the product of a low rank matrix $\boldsymbol{U}$, representing the primary directions of channel variance, and the channel factor vector x, a normally distributed random vector.

Using JFA, the speaker training is performed by calculating the full speaker dependent GMM supervector and discarding the channel dependent component. In the verification phase the

channel-dependent component can be estimated from the test utterances and the entire supervector can be scored using the linear dot-product approach [26].

2.3.3.5.2 I-vectors

It has been suggested that the channel space of JFA contains information that can be used to distinguish speakers [27]. Thus, discarding the channel dependent subspace of the JFA results in a loss of information and a degradation of the system overall performance. In an attempt to overcome this issue, a new approach to front-end factor analysis has been proposed, termed *i-vectors*.

Unlike in JFA where speaker and channel dependent subspaces are represented separately, i-vectors represent the GMM super-vector by a single *total-variability* space:

$$\boldsymbol{\mu} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}, \tag{23}$$

where $\boldsymbol{m}$ is the same UBM supervector used in the JFA approach and $\boldsymbol{T}$ is a low rank matrix representing the primary directions of variability across all training data. $\boldsymbol{w}$ are the total variability factors, represented by an independent normally-distributed random vector.

The classification can be done using cosine similarity scores (CSS), which operates by comparing the angles between a test i-vector $\boldsymbol{w}_{test}$ and a target i-vector $\boldsymbol{w}_{target}$:

$$\text{score}(\boldsymbol{w}_{target}, \boldsymbol{w}_{test}) = \frac{\langle \boldsymbol{w}_{target}, \boldsymbol{w}_{test} \rangle}{\|\boldsymbol{w}_{target}\| \|\boldsymbol{w}_{test}\|}. \tag{24}$$

2.3.4 Score normalization

Normalizing the distributions of scores can produce further improvements in system performance [28] and be an effective technique to increase the system robustness. The following examples are the most common approaches for scores normalization, which are used to compensate inter-speaker variability:

Zero normalization (Z-norm) [21]: A normalization technique which uses a mean and variance estimation for distribution scaling. The advantage of Z-norm is that the estimation of the normalization parameters can be performed offline during training. It calculates the scores of the target speaker model λagainst a set of impostor speech utterances. Then the mean μ_λ and standard deviation σ_λ of these scores are estimated to normalize the target speaker score $S(X, \lambda)$ computed from each utterance X against this target model. The normalized score has the form:

$$S_{znorm}(X, \lambda) = \frac{S(X, \lambda) - \mu_\lambda}{\sigma_\lambda}, \tag{25}$$

Test normalization (T-norm) [28]: A technique used to normalize the target score relative to an impostor model ensemble. The T-norm parameters are estimated from scores of each utterance X against a set of impostor speaker models at test time. Then the mean μ_X and the

standard deviation σ_X of the impostor scores are used to adjust the normalized target speaker score:

$$S_{tnorm}(X,\lambda) = \frac{S(X,\lambda) - \mu_X}{\sigma_X}, \tag{26}$$

It is possible to combine Z- and T-norm into a single normalization.

2.3.5 Evaluation metrics

To evaluate the performance of an SV system, one must subject it to a set of trials.

Each trial is accepted or rejected, and (as in any other binary classification problem) there are two types of errors that can occur: miss (or false rejection), defined as an incorrectly rejected target trials; and false alarm, an incorrectly accepted impostor trial. The performance of a verification system is the measure of the trade-off between these two errors (which is usually adjusted by a decision threshold variable).

In an evaluation, N_t target trials and N_i impostor trials are conducted and error probabilities are estimated at the threshold θ as follows:

$$Pr(miss \mid \theta) = \frac{num.target\ trial\ scores < \theta}{N_t} \tag{27}$$

$$Pr(false\ alarm \mid \theta) = \frac{num.impostor\ trial\ scores > \theta}{N_i} \tag{28}$$

The typical plots that depict system performance are the receiver operator characteristic (ROC) curve and the detection error trade-off (DET) curve. Both plot the probability of false alarm vs. probability of miss. Fig. 2.8 shows an example of each of these two curves. By setting the operating point at $P(miss) = P(fa)$, the equal error rate (EER) condition is found, which is often quoted as a summary performance measure.

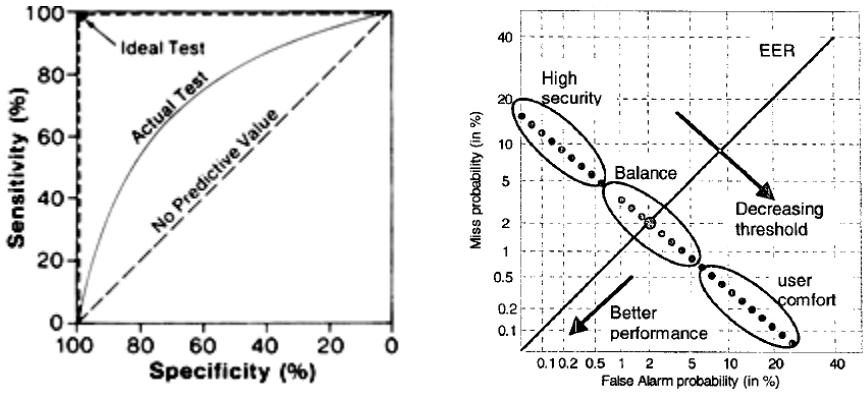

Fig. 2.8 Example of ROC (left) and DET (right) curves

Other measure of performance is the decision cost function (DCF) defined as:

$$DCF(\theta) = C(miss)Pr(target)P(miss|\theta) + C(fa)Pr(impostor)Pr(fa|\theta), \tag{29}$$

where $C(.)$ are cost parameters. This metric is useful for systems with specific costs for each type of error.

2.4 Voice conversion

VC is the adaptation of the characteristics of a source speaker to those of a target speaker in such way that an utterance of the source speaker can be transformed and sound like it was spoken by target speaker, while keeping the original language content [3].

It is important to make a distinction with related problems: *Voice transformation* is a general problem that encompasses all the methods that modify any features of a voice signal; *Voice Morphing* is a particular case of voice transformation where two voices are blended to form a third one, different from the original voices. Usually both speakers speak the same thing synchronously; *VC* systems take into account both *timbre* and *prosody* of the source and target speakers. Although these qualities are easy to be recognized by humans, they are hard to define in general terms. In the context of VC, timbral features are mainly associated with the dynamic spectral envelope of the voice signal, whereas the prosodic features are related to pitch/energy contours and rhythmic distributions of phonemes [29].

VC systems have many applications, for instance in all systems that make use of pre-recorded speech, such as voice mailboxes or elaborate text-to-speech (TTS) synthesizers. In these cases VC would be an efficient way to personalize the system without the need to retrain a full TTS system for each target speaker. In terms of medical applications, VC system would also improve the quality of life of persons who suffer from some form of voice pathology or who have had some form of surgery that renders them speech impaired, restoring their ability to communicate without losing their identity.VC also has applications in the field of machine translation, called *cross lingual voice conversion,* where the goal is to translate utterances from one language to another whilst preserving the characteristics of the original speaker in the translated speech. [30] The usefulness and comfort of those applications rely heavily on the quality and natural sounding of the speech generated from the VC system and this has served as a strong motivation for all the advances done in this field.

2.4.1 Voice conversion systems

In order to achieve full VC, voice source, vocal-tract and prosodic characteristics should be transformed. However, in practice, the actual features that are converted vary depending on the application. In the case of identity conversion, the mostly transformed features include spectral envelopes and average pitch and duration. These features are enough to provide a high degree of speaker discrimination by humans when the transformation is performed between two speakers with similar prosodic characteristics and patterns. In the case of emotion conversion, the most commonly transformed features are pitch and duration contours, which are the features that make a difference in this case. In the case of speech repair applications, the spectrum of modifiable features is much broader as the deviant features that need to be repaired may be of any kind.

Up to date, many different approaches have been tried to perform VC. Some using statistical techniques like GMMs [31], HMMs [32], unit selection [33] or principal component analysis (PCA)[34]; others using cognitive techniques such as artificial neural networks (ANN) [35]; even signal processing techniques like vector quantization [36] and frequency warping [37]. In the

following sections only explore the dominant trends in state-of-the-art approaches of VC, the GMM-based, and also unit selection-based, which has gained more popularity recently.

2.4.1.1 Gaussian mixture model

One of the most popular methods to use in a VC system is based on the estimation of the transformation function between the source and target speaker feature spaces using GMMs, as proposed by Stylianou [38]. When it was first introduced, it assumed parallel databases and featured several originalities comparing to other contemporary techniques. Among them were soft clustering, which meant that this model allowed continuous and smooth classification indexes, reducing artifacts generated by unnatural discontinuities; incremental learning, in order to minimize the influence of local errors; and continuous transform, which meant that each class was considered as a complete cluster rather than a single vector, reducing unwanted spectral distortions previously observed. Supported by all these advances, GMM training became a popular strategy at that time and currently it is still a baseline method to which novel approaches are usually compared to.

GMM-based VC can be achieved with both parallel and non-parallel databases [39]. In the former case it is assumed that the there are two sets of training data, one from the source speaker and another from the target speaker that contain the same set of utterances. These sets comprise N frames of spectral vectors $\boldsymbol{X} = [\boldsymbol{x}_1{}', \boldsymbol{x}_2{}', \dots, \boldsymbol{x}_N{}']'$, where $\boldsymbol{x}_n \in \mathbb{R}^d$ of the source speaker, and M frames of spectral vectors $\boldsymbol{Y} = [\boldsymbol{y}_1{}', \boldsymbol{y}_2{}', \dots, \boldsymbol{y}_M{}']'$ where $\boldsymbol{y}_m \in \mathbb{R}^d$ of the target speaker. The alignment between spectral frames is usually achieved with dynamic time warping algorithm (DTW) algorithm, which calculates an optimal match between the two sequences. In the latter case there is no correspondence between the utterances in the source and target speaker training data sets. In this case, the alternative is to use algorithms like [40], which find phonetically correspondent frames between speakers. To do so, usually there is an automatic method to perform segmentation of phonetic classes, followed by a mapping of the corresponding classes for the source and target speaker.

The GMM based VC as proposed in [38] is as follows. Firstly, an estimation of density model of the joint speaker feature spaces is performed, using the available training data:

Given a set of aligned training data comprised of N frames of spectral vectors $\boldsymbol{X} = [\boldsymbol{x}_1{}', \boldsymbol{x}_2{}', \dots, \boldsymbol{x}_N{}']'$, where $\boldsymbol{x}_n \in \mathbb{R}^d$ of the source speaker, and M frames of spectral vectors $\boldsymbol{Y} = [\boldsymbol{y}_1{}', \boldsymbol{y}_2{}', \dots, \boldsymbol{y}_M{}']'$ where $\boldsymbol{y}_m \in \mathbb{R}^d$ of the target speaker, the source and target feature vectors are combined in feature vector pairs $\boldsymbol{Z} = [\boldsymbol{z}_1{}', \boldsymbol{z}_2{}', \dots, \boldsymbol{z}_T{}']'$, where $\boldsymbol{z}_t{}' = [\boldsymbol{x}_n{}', \boldsymbol{y}_m{}']' \in \mathbb{R}^{2d}$. Then the joint probability density of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is modeled by a GMM:

$$P(\boldsymbol{X}, \boldsymbol{Y}) = P(\boldsymbol{Z}) = \sum_{i=1}^{M} \alpha_i^{(z)} \mathcal{N}\left(\boldsymbol{z} \middle| \boldsymbol{\mu}_i^{(z)}; \boldsymbol{\Sigma}_i^{(z)}\right), \tag{30}$$

where $\boldsymbol{\mu}_i^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_i^{(x)} \\ \boldsymbol{\mu}_i^{(y)} \end{bmatrix}$ and $\boldsymbol{\Sigma}_i^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{(xx)} & \boldsymbol{\Sigma}_i^{(xy)} \\ \boldsymbol{\Sigma}_i^{(yx)} & \boldsymbol{\Sigma}_i^{(yy)} \end{bmatrix}$ are the mean and the covariance, respectively, of the M-variate normal distribution $\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_i^{(z)};\boldsymbol{\Sigma}_i^{(z)})$. For each mixture component i, $\alpha_i^{(z)}$ is its prior probabilities, respecting the condition $\sum_{i=1}^{M} \alpha_i^{(z)} = 1$.

In the same way as an SV system, the parameters $\lambda^{(z)} = \left\{\alpha_i^{(z)}, \boldsymbol{\mu}_i^{(z)}, \boldsymbol{\Sigma}_i^{(z)} | i = 1,2, \ldots, M\right\}$ are estimated using the EM algorithm [20].

The density model is adopted to formulate a parametric transformation function to predict the target speaker's feature vector $\hat{\boldsymbol{y}} = F(\boldsymbol{x})$, for a given a source speaker feature vector $\boldsymbol{x}$:

$$F(\boldsymbol{x}) = E(\boldsymbol{y}|\boldsymbol{x}) = \sum_{i=1}^{M} p_i(\boldsymbol{x}) \left(\boldsymbol{\mu}_i^{(z)} + \boldsymbol{\Sigma}_i^{(yx)} \left(\boldsymbol{\Sigma}_i^{(xy)} \right)^{-1} \left(\boldsymbol{x} - \boldsymbol{\mu}_i^{(z)} \right) \right), \tag{31}$$

where $p_i(\boldsymbol{x}) = \frac{\alpha_i \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_i^{\boldsymbol{x}};\boldsymbol{\Sigma}_i^{\boldsymbol{xx}})}{\sum_{k=1}^{L} \alpha_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k^{\boldsymbol{x}};\boldsymbol{\Sigma}_k^{\boldsymbol{xx}})}$ is the posterior probability of the source vector belonging to the i^{th} mixture component.

The conversion function parameters are obtained by least squares optimization using the training data in order to minimize the total squared conversion error between the converted and target data. The error is given by:

$$\epsilon = \sum_{t=1}^{n} |y_t - F(y_t)|^2 \tag{32}$$

The framework of a GMM-based VC system is depicted in Fig. 2.9.

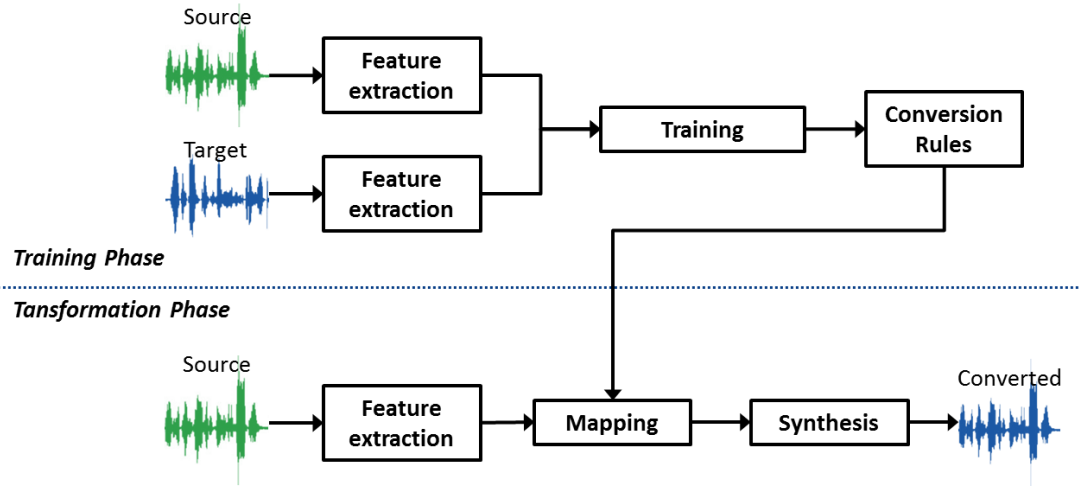

Fig. 2.9 Basic structure of a GMM-based VC system including training and transformation phases

2.4.1.2 Unit selection

Unit selection first arose in text-to-speech applications to generate natural-sounding synthesized speech [41]. The idea behind unit selection was to concatenate units (p. e. diphones, phones or ever frames) selected from annotated single speaker databases to synthesize a given utterance. Larger databases were usually preferred as there were more examples of the units with varied prosodic and spectral characteristics.

Later on, the concept of unit selection was imported to VC and is presently used as a strategy to achieve an optimal conversion between phonetically equivalent source and target units in VC systems with either parallel or non-parallel databases [42][43].

A unit selection-based VC system generally has two cost functions: given a unit database U and a target database T, the target cost $C^t(u_m, t_m)$ is an estimate of the difference between the database unit u_m and the target t_m which it is supposed to represent; and the concatenation cost $C^c(u_{m-1}, u_m)$ is an estimate of the quality of a joint between consecutive units u_{m-1} and u_m.

In terms of unit length, the most suitable for VC is a single frame, since it allows the independence of additional linguistic information about the processed speech. Hence, the cost functions can be defined by interpreting the feature vectors as database units i.e. $t \coloneqq x$ and $u \coloneqq y$.

To determine the target vector sequence $\tilde{\boldsymbol{y}}_1^M$ best fitting of the source sequence $\boldsymbol{x}_1^M$, one has to minimize the sum of the target and concatenation costs applied to an arbitrary sequence of vectors taken from a target sequence $\boldsymbol{y}_1^N$. Since all the compared units have the same structures (they are feature vectors) and have the same dimensionality, the cost functions are represented by Euclidean distances:

$$\tilde{\boldsymbol{y}}_1^M = \arg\min_{y_1^M} \sum_{m=1}^{M} \{\alpha S(\boldsymbol{y}_m - \boldsymbol{x}_m) + (1-\alpha) S(\boldsymbol{y}_{m-1} - \boldsymbol{y}_m)\}, \tag{33}$$

where α is a parameter to adjust the trade-off between fitting the accuracy of source and target sequence and the spectral continuity criterion.

Fig. 2.10 shows the typical framework for a unit selection-based VC system.

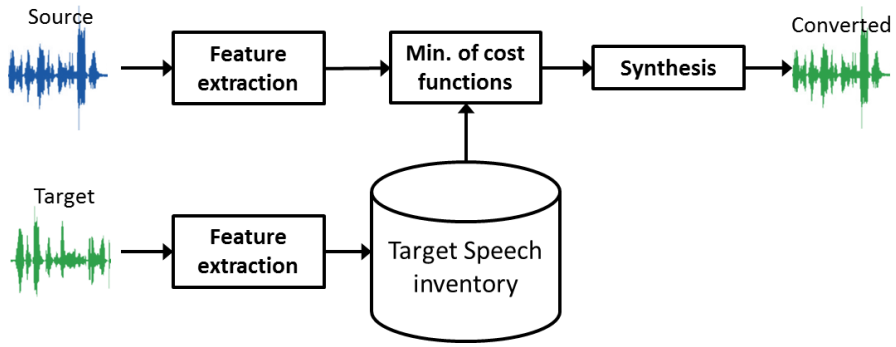

Fig. 2.10 Basic structure of a unit selection-based VC system including training and transformation phases

2.4.2 Evaluation metrics

A successful VC system must be capable of producing *natural*, *intelligible* and *identifiable* speech. The first quality is related to how human-like the converted speech sounds; the second is the easiness with which the converted words can be correctly identified; the last is how recognizable the individuality of the speech is. Different methods have been proposed to measure these qualities, some are *objective*, which are directly computed from the audio data and are considered cheap and fast; others are *subjective,* which are based on the opinions expressed by humans when faced against listening evaluation tests.

2.4.2.1 Objective measures

Distance measures are the most commonly used method to provide objective scores for the performance of VC systems. From the possible measures it is important to choose those that convey meaning in terms of the likeliness of the converted speech to the target speech. Among the significant measures, a popular one is the spectral distortion between the transformed and target speech which is computed as follows:

$$R = \frac{SD(trans, tgt)}{SD(src, tgt)}, \tag{34}$$

where R is the normalized distance, $SD(trans, tgt)$ is the spectral distortion between the transformed and the target speaker utterances and $SD(trans, tgt)$ is the spectral distortion between the source and target speaker utterances. The distortion can be computed, for instance, from log-likelihood ratio, LPC parameter, cepstral distance or weighted slope spectral distance measures.

Alternatively the comparison of the performance of different types of conversion functions can be computed using warped root mean square (RMS) log-spectral distortion [38].

Mel Cepstral Distortion (MCD) is another popular measure that stands out for having a correlation with the results of subjective tests [44]. Thus, MCD is usually used to measure the quality of the transformation [45]. MCD is related to the vocal characteristics and hence it is an important measure to check the performance of the mapping obtained by the source and target speakers. Essentially MCD is a weighted Euclidean distance defined as:

$$MCD = \frac{10}{\ln 10}\sqrt{2\sum_{k=1}^{25}(c_k^e - c_k^t)^2}, \tag{35}$$

where c_k^t and c_k^e denote the target and estimated MFCCs, respectively.

2.4.2.2 Subjective measures

The main advantage of using subjective measures is that they are directly related to the human perception, which is typically the standard for judging the quality of transformed speech. On the other hand they have the disadvantages of being expensive, time-consuming and difficult to interpret.

Among the most popular identity tests there is mean opinion score (MOS), which is a test used to evaluate the naturalness and intelligibility of the converted speech. In this test the participants are asked to rank the transformed speech in terms of either one or both of the aforementioned qualities. In MOS the participants are not exposed to speech from the target speaker so there is no possible measure for identifiability.

Another popular measure is the similarity test, in which the converted speech is compared to the target speech and graded from 1 to 5 in terms of similarity. With this test it is possible to obtain a measure of identifiability. Alternatively, ABX listening tests can also be performed to obtain a measure of similarity. In this test the participant is asked if the converted utterance X is more similar to the corresponding source or target utterance (either A or B).

3 Vulnerability of speaker verification systems against converted speech attacks

3.1 Introduction

For SV systems, security against impostor speakers is one of the most important problems. Over the course of the years many approaches to reduce false acceptances as well as misses have been investigated. Some of these have been very successful, keeping the error rates as low as 2% or 3%, or even less. However, the apparent robustness of SV systems falls short when the SV system is faced against converted spoofing data.

Until recently, developing synthetic of converted speech for a target speaker was a difficult task, but from the moment that VC systems became sophisticated and reliable enough they have posed a serious a problem for SV systems. The first study to address this issue reported FAR as high as 70% for synthetic speech against an SV trained only with natural speech [47], as it is the usual. Other studies performed more recently using more sophisticated SV systems also report the same kind of results, regardless of the SV systems having become more sophisticated.

The goal of this chapter was to confirm the reported vulnerability of a state-of-the-art SV system against converted speech. As such, a state-of-the-art SV system was built and its performance evaluated against a baseline corpus of natural speech. Then the same SV system was tested against two corpora of converted speech, each achieved using a different method. The performance of the SV system in these three conditions is compared.

3.2 Corpora

The experiments performed in this thesis to evaluate the performance of an SV system used several speech corpora for training and testing purposes which will be briefly described in this section. Both natural and converted speech corpora were used.

The natural corpus used in this thesis is a subset of the NIST SRE2006 corpus. The choice of this database came from its popularity in speaker detection tasks and the fact that it's publicly available. This allows it to be used by the whole scientific community and to make the performance of the developed systems more comparable. The NIST SRE2006 corpus is comprised partially of conversational telephone speech data collected from the Mixer Corpus by the Linguistic Data Consortium using the "Fishboard" platform, and partially data from "multi-channel" auxiliary microphones collected simultaneously. The data includes silence segments as it is not edited. The corpus features mostly data in English but also contain some speech in other languages. The NIST SRE2006 corpus is divided in five training conditions and four testing conditions.

The subsets of NIST SRE2006 corpus used in this thesis corresponds to one of the training conditions and one of the testing conditions, which together account for what is designated by

NIST as the "core task". Each train and test data file of the core task contains one five minute two-channel (4-wire) conversation involving one speaker on each side (1conv4w). The training data contains a total of 692 files, of which 462 belong to female speakers and 230 to male speakers. The testing data contains 504 unique speakers, where 298 are females and 206 are males. The test trials account for a total of 3938 genuine trials, of which 2349 belong to females and 1692 belong to males; and 2782 impostor trials, of those 1636 belong to female and 1146 belong to males. Out of the available trials, 1458 target and 2108 impostor trials were randomly selected for testing purposes, as summarized on Table 3.2.

The spoofing corpora available were provided by Zhizheng Wu and Haizhou Li of the Nanyang Technological University, Singapore who designed and created them using two different VC methods: GMM and unit-selection, as described in sections 2.4.1.1 and 2.4.1.2, respectively.

To ensure the comparability of the two spoofing corpora, both methods used the same source and target speakers data. The source speakers' data consisted of two training conditions of the NIST SRE2006, particularly of the 3conv4w and 8conv4w. The target speakers were randomly selected speaker models from the training condition 1conv4w of the NISTSRE2006. As such, the converted speech corpora feature converted impostor trials corresponding to the same speakers of the natural corpora available. The number of available trials used for testing purposes is 2164 for GMM-based converted speech and 2196 for unit selection-based converted speech.

Natural Corpora		**Converted Corpora**	
SRE2006 1conv4w train data	*SRE2006 1conv4w test data*	*GMM-based converted speech*	*Unit selection-based converted speech*
782	3647	2747	2748

Table 3.1 Corpora avaliable for the various experiments carried out in this study

Corpora	**Trials**	
	Target	**Impostor**
SRE2006 1conv4w test data	1458	2108
GMM-based converted speech	0	2164
Unit selection-based converted speech	0	2196

Table 3.2 Trials available for each corpora used for testing purposes

3.3 Baseline speaker verification system

In order to evaluate the performance of a state-of-the-art SV system against spoofing attacks, a baseline SV system was designed and implemented based on i-vectors, as explained in Section 2.3.3.5.2. The system used in-house technology, and the parameters were set to the values that provided the best results in extensive SV tests (prior to this thesis). The SV system is gender-dependent as it is commonly assumed in speaker recognition tasks that the speaker gender is known.

The features extracted from the training data of the natural corpora to build the model were 13 MFCCs, including the 0^{th} coefficient, and their respective velocities and acceleration parameters, totaling a 39 dimensional feature vector. Cepstral mean and variance normalization (MVN) is applied in a per segment basis. Finally, low-energy frames detected with the alignment generated by a simple bi-Gaussian model of the log energy distribution computed for each speech segment are removed.

A GMM-UBM composed by 1024 mixture components was trained using the training data from the 1conv4w condition of NIST SRE 2004 and 2005 corpora using the same MFCC features. The same VAD algorithm based on the bi-Gaussian model was used to eliminate the feature vectors corresponding to silence.

The total variability factor matrix, T, was estimated according to [49]. The dimension of the total variability sub-space was fixed to 400. Zero and first-order sufficient statistics of the training natural corpora described in Table 3.1 were used for training T. Ten EM iterations were applied. For each iteration first ML estimation updates were applied, followed by minimum divergence update. The covariance matrix was not updated in any of the EM iterations.

No channel compensation or probabilistic linear discriminant analysis (PLDA) were applied as post-processing techniques, although that is a usual practice, which makes this a simple SV system. This should not be seen as an issue as the purpose of this SV system is only to demonstrate the vulnerability of a generic SV system. It would not make sense to spend too much effort implementing a very complex SV system for such a purpose.

Finally, the verification score is obtained by cosine similarity, as indicated in Eq. 24, between the target speaker i-vector and the test segment i-vector.

3.4 Experimental results

The first experiment to establish the baseline performance of the SV system was to test it against natural same-gender speakers' data. The trials set consisted of 842 female and 616 male natural target speaker trials and 1091 female and 1017 male natural impostor trials. The performance of the SV system is evaluated using the EER, a standard metric for speaker detection tasks.

The SV system achieved a performance of 8.9% EER for female trials and 9.9% for male trials. Other reports, such as [6] evaluate this same task (core task of NIST SRE2006) and achieved performances around the 3% EER. These reports used SV systems based on technologies other than i-vectors. In [27], the core task of NIST SRE2008, a similar task yet not the same as in this experiment, was evaluated using i-vector based SV systems and achieved a performance of around 2% EER. This shows the potential of using i-vectors for SV tasks. From these, it can only be assumed that the justification for a poorer baseline performance in this study is a consequence of a poor pre- or post-processing, and not because of the i-vector method itself. Examples of pre- or post-processing methods that could have contributed for the degradation of the performance include the VAD algorithm and the intersession compensation that was not

performed. However the performance of the system is still good in order to be used as a baseline reference.

To evaluate the vulnerability of the SV system against converted speech, two more experiments were carried out, one for each of the available spoofing corpora, in order to understand if different VC approaches affected the performance of the SV system differently.

For the GMM-based spoofing corpus case, a new set of trials was created by merging the set of natural trials used previously with 1314 female and 850 male GMM-based converted impostor trials. The SV system was tested against these new sets of trials and achieved a performance of 19.0% EER for female trials and 21.0% for the male trials. This corresponds to a relative degradation of over 100% of the EER of the SV system performance. However, as only the number of impostor trials was manipulated, the spoofing attack should only affect the false acceptance rate, while the miss rate remained constant, which makes the EER a misleading metric for this task. A more meaningful metric would be to set the decision threshold using the EER point of the baseline performance and use the miss rate and FAR as the performance metrics. This would simulate what is done in a real life SV application, where a development set of data is used to determine the decision threshold that is subsequently used in trials. After doing so, the miss rate remained at 8.9% for female trials and 9.9% for male trials, as the number of target trials was the same as in the previous experiment. The FAR increased to 33.0% for female trials and 35.2% for male trials. Out of the total number of converted impostor trials, 52.9% of the female trials and 63.5% of the male trials were falsely accepted.

The same experiment was conducted for the unit selection-based spoofing corpus. The baseline sets of trials were augmented with 1421 female and 775 male unit selection-based converted impostor trials. The EER of the performance of the system for this experiment was 19.6% for female trials and 22.8% for male trials. Once again the relative EER degradation is over 100%. After setting the decision threshold to the EER point of the baseline trials, the FAR increased to 38.3% for female trials and to 39.3% for male trials. The miss rate remained the same as in the baseline, as mentioned. Out of all the unit selection converted trials, 60.7% of the female ones and 75.7% of the male ones were misdassified.

Tables 3.3 and 3.4 summarize the gender-depended performance achieved respectively by the female and male SV system in terms of miss rate and FAR for the baseline set of trials, the sets of trials with GMM-based converted speech and the sets of trials with unit selection-based converted speech. Additionally, Fig. 3.1 and Fig. 3.2 show the DET curves for the SV system performance for each gender and for the same three sets of trials.

Test data	Miss %	False acceptance %	Converted trials misclassifications %
Baseline (no converted speech)	8.9	8.9	-
Natural and GMM-based converted speech	8.9	33.0	52.9
Natural and unit selection-based converted speech	8.9	38.3	60.7

Table 3.3 Performance of the female SV system in miss rate, FAR and converted trials misclassification rate against natural and converted speech

Test data	Miss %	False acceptance %	Converted trials misclassifications %
Baseline (no converted speech)	9.9	9.9	-
Natural and GMM-based converted speech	9.9	35.2	63.5
Natural and unit selection-based converted speech	9.9	39.3	75.7

Table 3.4 Performance of the male SV system in miss rate, FAR and converted trials misclassification rate against natural and converted speech

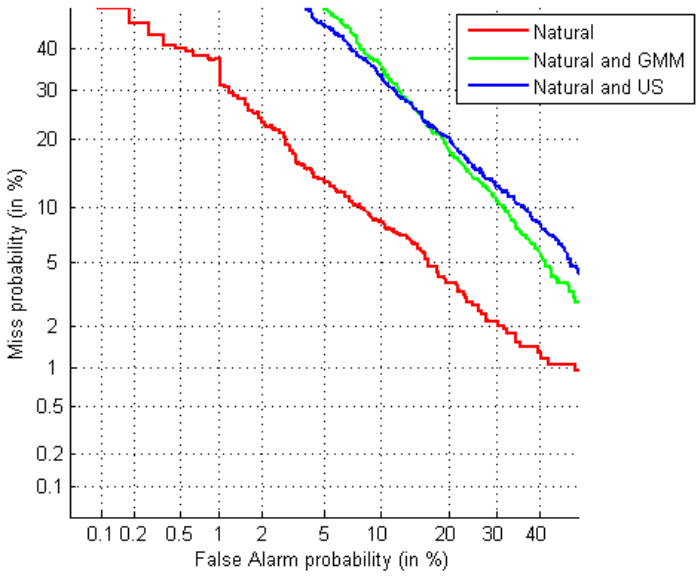

Fig. 3.1 DET curve for the performance of the female SV system against natural and converted data

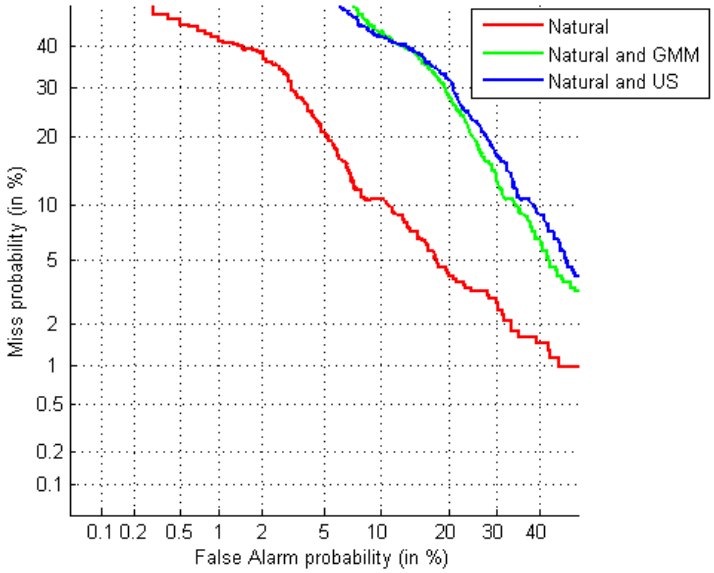

Fig. 3.2 DET curve for the performance of the male SV system against natural and converted data

Reporting the performance results of the systems in duplicate in order to cover the results from both genders makes the reading of this document heavier. In order to avoid so, the results presented from here onwards will feature a pooling of the female and male results into a single one. The results from Tables 3.3 and 3.4 and Fig. 3.1 and 3.2 are shown combined in Table 3.5 and Fig. 3.3.

It should be noted that, although the results are presented without specifying the gender, all of the performed experiments featured separate female and male models and trial sets before the pooling.

In terms of the results obtained in the evaluation of the vulnerability of the SV system against converted data, it is possible to observe from Table 3.5 and Fig. 3.3 that the FAR of both spoofing trial sets increases beyond what is considered acceptable for a state-of-the-art SV system, which is in accordance with previous studies. Additionally, the performance of the SV system when under the spoofing attacks was worse for the unit selection-based spoofing corpus.

The misclassification rate of converted speech trials showed alarmingly high scores proving the complete vulnerability of a state-of-the-art SV system without anti-spoofing mechanisms to a converted speech spoofing attack. In order to address this security issue, and make the performance of the SV system return to acceptable values, it is necessary to include some sort of protection in the system.

Test data	Miss %	False acceptance %	Converted trials misclassifications %
Baseline (no converted speech)	9.4	9.4	-
Natural and GMM-based converted speech	9.4	33.7	56.6
Natural and unit selection-based converted speech	9.4	38.5	65.6

Table 3.5 Pooled performance of the female and male SV system in miss rate, FAR and converted trials misclassification rate against natural and converted speech

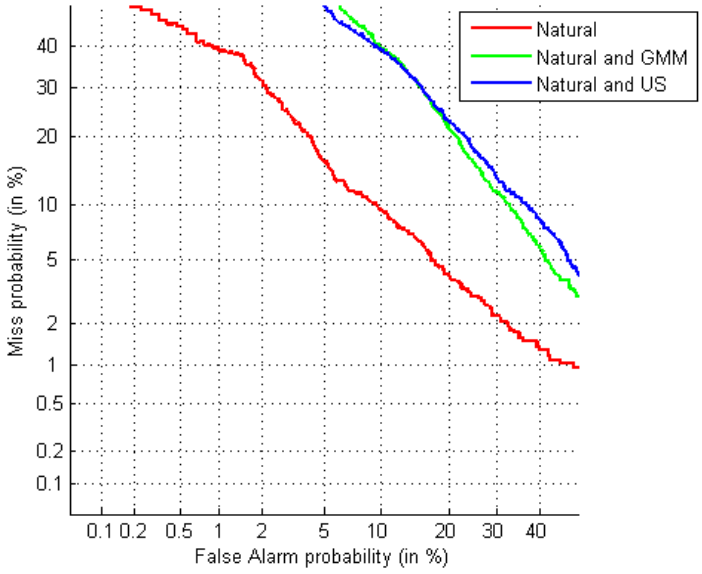

Fig. 3.3 DET curve for the pooled performance of the female and male SV systems against natural and converted data

4 Anti-spoofing

4.1 Introduction

Spoofing is any attack that is performed by an impostor that intentionally tries to fool a verification system. This type of attacks is common in biometric systems (using the face, voice or fingerprint recognition) and on internet authentication [50]. In the particular case of speech related applications, spoofing attacks may come in the form of a voice mimic (a person trying to impersonate the target speaker) [51], play back recordings [52], and the most recent and sophisticated form, through VC [53] [54].

As seen, SV is the process of confirming a claim of identity of a user based on his speech samples. The outcome of this claim is either an acceptance of the claimed identity or a rejection. On the other hand, VC is the modification of a source speaker's voice to sound like the voice of a target speaker. Hence, a VC technique can be used to modify a voice of an impostor to make it sound like the voice of a claimed speaker, allowing it to perform a spoofing attack against an SV system. In section 3.3 it was shown the actual vulnerability of a state-of-the-art SV system against a this type of attack.

In response to such potential threads, numerous anti-spoofing efforts, meaning efforts to evaluate the vulnerability of speaker verification systems against spoofing attacks and to develop efficient countermeasures, have been performed. Most of the countermeasures are inspired by the limitations of VC systems. Hence, an overview of the limitations of both SV and VC systems is essential to understand which are the anti-spoofing approaches that seem most promising. After that, some contextualization in terms of reported anti-spoofing mechanisms is given.

4.2 Limitations of speaker verification systems

The advances in biometrics and their increasing use in everyday applications are one of the leading factors that have stimulated the improvements in SV systems. Overtime they have become more complex and robust, capable of solving more difficult classification tasks. However SV systems are still vulnerable because of several overlooked issues:

- *Phase information* [55]: Most of the features used in SV (and speaker identification) systems are derived from magnitude spectrum, the most common being the MFCCs, like described in section 2.2.1. These features however, ignore the phase information of the Fourier spectrum, as the phase spectrum is discarded in the feature extraction process. While it is true that most of the perceptual information about speech lies in the magnitude spectrum, it has been shown that phase spectrum also carries important information about voice identity [56], thus using this information to build a speaker models will make them more robust.
- *High level cues*: It is well known that the low-level perceptual cues carry information about the acoustic characteristics of speech and high level cues carry linguistic aspects

of speech. The low-level cues can be captured by the typical features that are extracted at frame level, like the MFCCs, hence, to some extent the speaker is characterized by them. However, the high-level cues are left out as they are related to longer speech patterns such as intonation, stress and timing [57]. The ability for humans to distinguish between speakers drastically improves when the listener is familiar to the speaker. This is due to listener being able to identify the speaker characteristic idiosyncrasies, either consciously or unconsciously [57], proving that the high-level cues also carry important information about speaker identity. Further improvements of SV systems surely include developing more sophisticated methods of feature extraction that explore these long-term features, and by including these new features in the speaker models [22][58]. Additionally, high-level features may also help improve robustness since they should be less susceptible to channel effects [22].

4.3 Limitations of voice conversion systems

VC systems have undergone huge developments in the last decades, and state-of-the-art systems, like those briefly described in section 2.4 have shown better results in both objective and subjective evaluations than their predecessor techniques. However, VC is still not perfected. In fact, it still has some important limitations that come from the some overlooked phenomena:

- *Phase information* [59]: likewise to SV systems, VC systems use features that are derived from the magnitude spectrum of the speech signal. These features are then transformed and passed to the speech synthesizer filter which reconstructs the speech using only information from magnitude parameters, ignoring the phase information. As a result the original/natural phase information is not kept in the reconstructed waveform. This phenomenon can be detected objectively by analyzing the phase spectrum of the converted speech and searching for artifacts in it and subjectively through listening tests.
- *High level cues* [30]: Once again, the modeling techniques used in VC are similar to the ones used in SV systems. In both cases the model is built using features extracted at frame level. These features only contain information about low-level cues. In VC systems this issue has impacts in terms of intonation modulation (among others), which is related to high-level cues. Consequently the converted speech will not have the target speaker natural intonation because it has not been converted.
- *Quality issues* [31] [60] [61]: Although converted speech already sounds reasonably natural, many issues have been reported in terms of lack of quality of the converted speech, namely hissing noise, ringing tones, clicks and also timbral aspects that may be described as unnatural or synthetic voice. These include large pitch shifts without formant correction which might degrade quality and even intelligibility of the converted speech. Such issues are very easily detected in subjective tests.

4.4 Anti-spoofing mechanisms

State-of-the-art SV systems do not have any particular way of protecting themselves against spoofing attacks. Hence, when an SV system is systematically faced against converted speech, it tends to increase its FAR, thus degrading its performance [6]. The increase of the FAR is very inconvenient when the application needs to be robust. In order to overcome this issue and decrease the FAR, it is useful to introduce new modules in SV systems that detect converted speech [7]. The state-of-the-art techniques used in these modules are inspired by the limitations of both SV and VC systems in such way that they exploit the imperfections of converted speech. These converted speech detector modules can be included as a post-processing module, after the SV system itself. Their input should be only the test utterances that were accepted by the system. The output should be a binary decision between converted or natural speech.

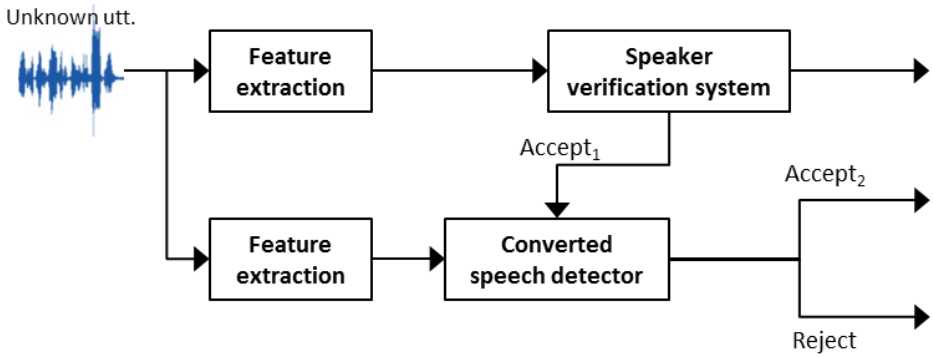

Fig. 4.1 Basic structure of an SV system with a converted speech detector as an anti-spoofing mechanism

4.4.1 Extracting information from phase spectrum

As mentioned, one of the major limitations of state-of-the-art VC systems is that they discard the phase spectrum when extracting meaningful features to convert. Thus the natural speech phase information is lost during the reconstruction of the signal, creating artifacts in the phase spectrum of the converted speech. A possible approach to create a converted speech detector that spots these artifacts in the phase spectrum was studied in [7].

In order to extract features derived directly from the phase spectrum of a speech signal, it is necessary to compute the unwrapped phase [62]. An alternative that is computationally simpler is using the group delay function (GDF) [62], which has the additional advantage of reducing the effects of noise.

The GDF is a measure of non-linearity of the phase spectrum [64] and is defined as the negative derivative of the phase spectrum with respect to the frequency. For a given signal let:

$$X(\omega) = |X(\omega)|e^{i\phi(\omega)} \tag{36}$$

$$\log\big(X(\omega)\big) = \log(|X(\omega)|) + j\phi(\omega) \tag{37}$$

$$\tau(\omega) = -\frac{d\phi(\omega)}{d\omega}. \tag{38}$$

The GDF, $\tau(\omega)$, can be computed directly from the speech signal:

$$\tau(\omega) = -\left(\frac{d(\log(X(\omega)))}{d\omega}\right)_I = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|X(\omega)|^2}, \quad (39)$$

where $X(\omega)$ and $Y(\omega)$ are the STFT of $x(n)$ and $nx(n)$, $X_R(\omega)$, $X_I(\omega)$, $Y_R(\omega)$ and $Y_I(\omega)$ are the real and imaginary part of $X(\omega)$ and $Y(\omega)$, respectively.

It is known that the GDF yields a spiky nature, primarily caused by pitch peaks, noise, and window effects. In this sense, sometimes it is more convenient to adopt a modification of the GDF that suppresses these effects. The modified group delay function (MGDF) as proposed in [66] is achieved by using a smoothed version of the power spectrum. The MGDF can be reshaped by introducing two new parameters. The final MGDF is given by:

$$\tau_{\alpha,\gamma}(\omega) = sign. \left| \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{(S(\omega))^{2\gamma}} \right|^{\alpha}, \quad (40)$$

where $sign.$ is the sign of $\frac{X_R(\omega)Y_R(\omega)+Y_I(\omega)X_I(\omega)}{S(\omega)^2}$, $S(\omega)$ is the smoothed version of $|X(\omega)|$, which is obtained by cepstral smoothing, and α and γ are parameters that must be optimized to further reduce the spiky nature of the spectrum.

From the GDF or the MGDF it is possible to extract coefficients that represent the spectrum in an efficient way. For a given speech signal $x(n)$, the group delay cepstral coefficients (GDCCs) can be derived as follows:

1) Compute STFT $X(\omega)$ and $Y(\omega)$ of $x(n)$ and $nx(n)$, respectively;
2) Compute the GDF as in Eq. 39.
3) Apply DCT to the GDF,
4) Keep k cepstral coefficients, excluding the 0th coefficient.

The modified group delay cepstral coefficients (MGDCCs) can be derived in a similar fashion:

1) Compute STFT $X(\omega)$ and $Y(\omega)$ of $x(n)$ and $nx(n)$, respectively;
2) Compute the cepstraly smoothed spectrum of $|X(\omega)|$, $S(\omega)$;
3) Compute the reshaped MGDF as in Eq. 40;
4) Apply the DCT to the MGDF;
5) Keep k cepstral coefficients, excluding the 0th coefficient.

As mentioned the α and γ parameters will need to be optimized, which can be accomplished in a small development set. As a rule of thumb, the number of cepstral coefficients, k, to be kept can be the same as in MFCCs.

The GDCCs and MGDCCs can be used as feature vectors by any modeling technique and in particular by a converted speech detector.

4.4.2 Extracting information from feature trajectory

The most commonly used features in VC systems are derived from the power spectrogram in a frame-by-frame fashion. As a result, they do not capture most of the correlation between frames or the temporal characteristics of speech feature trajectories. As a consequence, the temporal information related to the high-level perceptual cues associated to them is lost. Converted speech achieved by any method that uses features without temporal information is may have temporal artifacts. As such, it is useful for a system concerned with the detection of converted speech to use features that help detect those artifacts.

It was recently proposed the use of modulation features to address this task. These features are said to capture enough temporal information by using features trajectories so that they allow an easy detection of temporal artifacts. The modulation features can be extracted from any spectrogram. We will contemplate modulation features extracted from the magnitude spectrum, called magnitude modulation (MM) features; and from the GDF or MGDF spectrum, which are called phase modulation (PM) features. These features contain information on the temporal evolution of the magnitude and phase spectrum of the speech signal, respectively. [8].

The MM coefficients can be extracted from a given speech signal, $x(t)$, by following the steps:

1. Compute the power spectrogram of $x(t)$;
1) Divide the power spectrogram into overlapping segments using n frames windows with m frames shift (n should be large enough that it captures temporal information, and m should be half of less of n);
2) Obtain c Mel-scale filter-bank coefficients from the spectrogram window, forming a $c \times n$ matrix;
3) Apply MVN to the trajectory of each filter bank to normalize the mean and variance to zero and one, respectively;
4) Apply a p-point FFT to the c normalized trajectories in order to compute the modulation spectrum from filter-bank energies;
5) Concatenate every modulation spectra on the spectrum to make up a modulation supervector with $\frac{cp}{2}$ dimensions (given the symmetrical nature of the FFT, only the first half of the coefficients needs to be kept);
6) Apply PCA to eliminate the highly correlated dimensions of the coefficients. The number of components to be kept varies with the problem. However, in most of the problems it is enough to keep the k components with the highest associated variance that account for 95% or so of the total variance.

The projected MM coefficients are k-dimensional can be used as feature vectors by any modeling technique. To compute the PM coefficients, the same steps should be followed only changing the spectrogram before applying the Mel-scale filter-bank to a GDF spectrogram or MGDF spectrogram, following either Eq. 39 and Eq. 40, respectively.

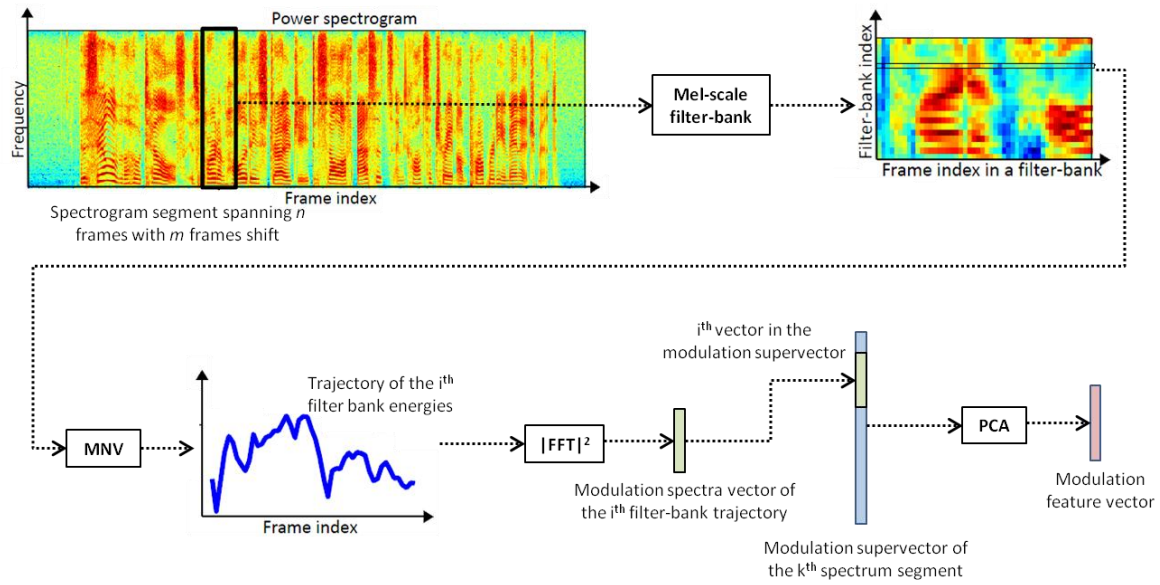

Fig. 4.2 Magnitude modulation feature extraction process as in [8].

4.4.3 Detecting differences in pair-wise distance between consecutive feature vectors

The features already described have a common goal of easing the distinction of natural and converted speech by exploiting the artifacts introduced during the conversion process. Some recent studies try to take a slightly different approach by searching for subtler differences in the particular cases of converted speech. In [10] the authors propose an approach to detect GMM-based converted speech which essentially tries to explore a natural consequence of using such model instead of the artifacts created through information loss.

These approach explores the fact that in GMM-based VC there is a phenomena caused by the conversion process that creates an inconsistent effect in the converted speech that is not present in natural speech: In VC, it is expected that consecutive feature vectors shift towards the same local maxima of the likelihood function of the target model, thus reducing the average distance between pair-wise feature vectors as shown in Fig. 4.3 and increasing the density of the features surrounding the local maxima.

With this in mind, it is possible to develop a speaker-dependent synthetic speech detector that uses pair-wise distances between consecutive feature vectors to discriminate between natural and converted speech [10]: Given a distribution in pair-wise distances between consecutive feature vectors in test data $s[n]$ and a distribution of the pair-wise distances between consecutive feature vectors in train data of the target speaker, the percentage of overlap between the two distributions forms a score that can be thresholded to classify $s[n]$ as natural or converted speech. Lower scores are an indicator of natural speech whereas higher scores indicate converted speech. Additionally, if the two distributions are normalized, the percentage of overlap lies between zero and unity.

Additionally, it is recommended that the countermeasure is implemented in the LPC space as it is the one that shows the biggest difference in the distance magnitudes between natural and converted speech (comparing to linear prediction cepstral coefficients and automatic SV feature space) [10].

The advantage of using such approach is that it is not feature-dependent. In this sense, as long as the converted speech was GMM-based it would be easier to recognize it regardless of the improvements in the amount of converted information. However, this detector is model-dependent, which means that it may be completely useless to detect voice converted with other methods that are not GMM-based. Further studies regarding this issue have never been performed.

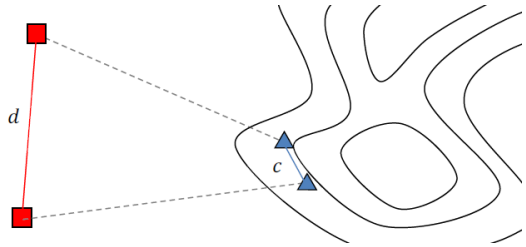

Fig. 4.3 Illustration of conversion of feature vectors in feature space showing the expected reduction of pair-wise distance caused by their shift towards a common local maxima of the target speaker model after Alegre et al.

5 Converted speech detectors as anti-spoofing mechanisms

5.1 Introduction

Following what is reported in the literature in terms of anti-spoofing, the dominant trend is to try to detect artifacts in the converted speech. These artifacts may be in the phase or magnitude spectrum of the signal and they may be a result of short- or long-term information loss. To perform a comprehensive study about converted speech detection it is necessary to design and implement converted speech detectors that use features that capture each of the four possible artifacts: short-term in the magnitude spectrum; short-term in the phase spectrum; long-term in the magnitude spectrum and long-term in the phase spectrum.

In this chapter the first experiments focus on reproducing state-of-the-art converted speech detectors using features meant to ease the detection of artifacts characteristic of converted speech. Four state-of-the-art converted speech detectors using GMMs as the modeling technique (as suggested in [7]) are implemented, each using a type of feature specifically chosen to detect one of the four types of possible artifacts.

After these experiments, new converted speech detectors are proposed, using SVM as the modeling technique and using a compact feature representation, that both aim at improving the detection accuracy of the system and at decreasing the amount of data necessary to train the models. Four new detectors are implemented, one for each type of spectral artifacts. The performance of the new detector is subsequently compared to the corresponding state-of-the-art one.

It has been previously reported in [8] that short- and long-term features may contain complementary information about the speech signal, as such, the possibility of fusing the scores of detectors using short- and long-term features is also featured in the experiments that were carried.

Finally the performance of all the implemented converted speech detectors (state-of-the-art, proposed, standalone and fused) is compared.

5.2 Corpora

All of the experiments carried out with the converted speech detectors used the same three corpora that were described in section 3.2. All the converted detectors were trained and tested with non-overlapping, randomly chosen subsets of each of the available corpora.

5.3 State-of–the-art converted speech detectors

Presently all the converted speech detectors reported in the literature are modeled by GMM de facto standard technique. The most successful converted speech detectors reported so far were introduced by [8] and are based on the detection of spectral artifacts. In the following section these detectors are reproduced.

5.3.1 Experimental results

The state-of-the-art converted speech detectors were implemented using the *de facto* standard GMM modeling technique, as described in section 2.3.3.2. The systems described in section 2.3.3.2 are meant to perform SV tasks, however they are easily adaptable into converted speech detectors: instead of training a model for each speaker, only two models are trained, one for natural speech and another for converted speech. The detectors hereafter described using shot-term features have 512 mixture components in the GMM and those using long-term features have 32 mixture components.

- *Detecting short-term magnitude spectrum artifacts:*

The GMM-based converted speech detector implemented to detect short-term artifacts in the magnitude spectrum used the standard MFCCs as features, using the extraction process described in section 2.2.1. The MFCCs were extracted in 20ms frames, updated every 10ms. Each feature vector had 12 MFCCs, no log-energy and was not augmented with temporal derivatives. A VAD algorithm based on decision-directed parameter estimation method for the likelihood ratio test [65] was used to eliminate the feature vectors corresponding to silence.

The model for natural speech, $\lambda_{natural}$, was trained with 300 randomly selected files of natural speech corpus described in section 3.2; the models, λ_{conv_GMM} and λ_{conv_US}, for GMM- and unit selection-based converted speech were trained with the same amount of data from the GMM- and unit selection-based spoofing corpus, respectively; a third model for converted speech, λ_{conv_mix}, was trained using mixed GMM- and unit selection-based converted speech data in equal amounts, totaling 300 files. The purpose of this last model is to provide a more generalized representation of converted speech, containing data of converted speech achieved with different methods.

The GMMs were tested against three sets of data: a subset of data from natural corpus not used for model training consisting of 2460 natural speech files; 2447 files of GMM-based converted speech not used for training and 2448 files of unit selection-based converted speech not used for training, from the GMM and unit selection spoofing corpus described in section 3.2, respectively.

Overall there were 9 test conditions, where each of the three test sets was tested against a pair of a natural and one of the three converted models.

The purpose of testing GMM-based converted data against a model of unit selection-based converted speech and vice-versa is to understand if the features are robust enough to identify converted speech unknown examples even if they were generated with unseen methods.

The decision, $\Lambda(X)$, was made according to Eq. 9 where the decision threshold was set to zero, making the trials with a score bigger than zero be predicted as natural speech and those with a score lower than zero as converted speech.

The results are presented in Table 5.1 and are measured in accuracy rate. One can observe that the MFCCs work reasonably well as features for converted speech detection, achieving accuracy rate of over 79% for every condition except when unit selection-based converted speech is compared with natural and GMM-based converted models. The low accuracy may be a consequence of the differences in short-term magnitude artifacts in GMM-based and unit selection-based converted speech. The most difficult type of speech to detect is unit selection-based converted speech.

The use of a mixed converted speech model did not led to any significant improvement of the system performance.

Log likelihood ratio	**Test data (Acc. %)**		
	Natural	**GMM**	**US**
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_GMM})$	87.0	89.5	20.6
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_US})$	85.4	95.3	89.3
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_mix})$	84.6	93.4	79.8

Table 5.1 Performance in accuracy rate of the GMM-based converted speech detector using MFCCs in 9 test conditions

- *Detecting short-term phase spectrum artifacts:*

Following the work of [6], a GMM-based converted speech detector was trained using MGDCCs as feature vectors in order to detect short-term artifacts in the phase spectrum. The features were extracted according to the process described in section 4.4.1, in 20ms frames, updated every 10ms. Each feature vector had 12 MGDCCs, no log-energy and no temporal derivatives. A VAD algorithm was used to eliminate the feature vectors corresponding to silence. During the computation of the MGDF, Eq. 40 contemplates two parameters that need to be defined, α and γ, additionally the smoothing of the power spectrum, $|X(\omega)|$, into $S(\omega)$ is achieved by cepstral smoothing. In this process the signal represented in the cepstral domains, filtered, and returned to the time domain. The filter in question, $w(\sigma)$, is a high pass window with a cutoff frequency of σ, making up a total of three parameters to be defined. The special case where Eq. 40 is used with $\alpha = 1, \gamma = 1$ and $\sigma = 0$ is equivalent to Eq. 39 , where the spectrum is not smoothed and the fine structure of the spectrum is not emphasized by α and γ . With this parameter configuration the resulting feature vectors are the GDCCs.

To understand the influence of each parameter in the information contained in the feature vectors, several configurations were contemplated:

- $\alpha = 1, \gamma = 1$ and $\sigma = 0$, to study the effects of not using the smoothing and reshaping parameters to modify the GDF spectrum;
- $\alpha = 0.3, \gamma = 0.9$ and $\sigma = 30$, a configuration proposed in [66] claimed to be the optimal modification for phoneme recognition tasks;
- $\alpha = 0.4, \gamma = 1.2$ and $\sigma = 30$, a configuration proposed in [6] claimed to be the optimal modification for converted speech detection tasks.

Using each parameterization the GDCCs, $MGDCCs_1$ and $MGDCCs_2$ were respectively extracted. It should be mentioned that other parameter configurations were tested but given to the poor results, they are not reported in this document.

The converted speech detectors were trained with, and tested against the same subsets of the available corpora as in the previous experiments. The results for the performance of the detectors using GDCCs, $MGDCCs_1$ and $MGDCCs_2$ (extracted using each of the three different parameter configurations) are presented in Tables 5.2, 5.3 and 5.4 respectively.

Log likelihood ratio	Test data (Acc. %)		
	Natural	GMM	US
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_GMM})$	91.0	89.4	52.1
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_US})$	96.1	86.8	74.0
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_mix})$	91.5	92.3	77.1

Table 5.2 Performance in accuracy rate of the GMM-based converted speech detector using GDCCs in 9 test conditions

Log likelihood ratio	Test data (Acc. %)		
	Natural	GMM	US
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_GMM})$	89.4	89.9	40.4
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_US})$	78.8	95.1	88.0
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_mix})$	85.6	93.2	76.8

Table 5.3 Performance in accuracy rate of the GMM-based converted speech detector using $MGDCCs_1$ in 9 test conditions

Log likelihood ratio	Test data (Acc. %)		
	Natural	GMM	US
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_GMM})$	95.6	94.6	14.4
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_US})$	73.3	98.4	92.3
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_mix})$	96.1	95.4	58.2

Table 5.4 Performance in accuracy rate of the GMM-based converted speech detector using $MGDCCs_2$ in 9 test conditions

The results in Tables 5.2, 5.3 and 5.4 show that, overall the three systems using the GDCCs, $MGDCCs_1$ and $MGDCCs_2$ features archived better accuracies than one using the MFCCs. On the other hand the inefficiency in detecting unit selection-based converted speech when comparing it with a GMM-based converted speech model shows the lack of robustness of the features when facing unseen examples of converted speech methods. However, that is not the case for the model of unit selection-based converted speech. The toughest type of speech to detect is still the unit selection based converted speech.

Out of the three features, the best performing one on average is the GDCC. If the performance of the system in the crossed method detection candidates (detecting converted speech

achieved with one method by comparing it with a model of another VC method) with unit selection based test data is excluded, then the best performing feature is the $MFCCs_2$.

- *Detecting long-term magnitude spectrum artifacts:*

The detection of long-term artifacts in the magnitude spectrum is done with a GMM-based detector using MM features. The MM features are extracted according to the process explained in section 4.4.2, where the spectrogram was computed using 20ms windows and 256-point FFT. The spectrogram was divided into segments of 50 windows, with a shift of 20 frames from the previous one. The Mel-scale filter bank contained 20 filters. After MVN, a 64-point FFT was applied to the 20 normalized trajectories. Hence the modulation supervector had 640 dimensions. PCA was applied to reduce dimensionality and only the 10 projected dimensions with highest associated variance were kept. These accounted for 97.1% of the total variability.

The converted detector was trained with MM features extracted from the same subsets already described and tested against the remaining data available, the same as previously. The results for the 9 test conditions evaluated are presented in Table 5.5.

Log likelihood ratio	**Test data (Acc. %)**		
	Natural	**GMM**	**US**
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_GMM})$	73.6	77.5	47.8
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_US})$	69.3	44.5	63.3
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_mix})$	62.7	74.3	75.0

Table 5.5 Performance in accuracy rate of the GMM-based converted speech detector using MM features in 9 test conditions

Comparing the results in Table 5.5 with those on Table 5.1 it is evident that the performance of this detector is poorer than the MFCCs performance. This may be for one of two reasons: either the MM features, as they were extracted, are not so good at capturing long-term artifacts as the MFCCs are ate capturing short-term ones, or there are in fact fewer artifacts in the magnitude spectrum related to long-term information loss than short-term information loss. The later seems a less reasonable assumption given that the VC methods that were used do not use features or include mechanisms that allow the conversion of long-term information.

The remaining tendencies that were previously mentioned for the previous detectors are maintained for this one.

- *Detecting long-term phase spectrum artifacts:*

The detection of long-term artifacts in the phase spectrum is made in a similar fashion than of those in the magnitude spectrum. As explained in section 4.4.2 the only difference from the extraction process used previously is that the spectrogram is replaced by a MGDF spectrogram. The experimental details are the same as in the previous experiments, hence the outcome of the process are 10-dimensional PM feature vectors. The 10 kept projected dimensions account for 97.4% of the total variability of the original dimensions.

In the experiments carried out to detect long-term artifacts in the phase spectrum, the same three parameters configurations that were used in 40 were used to compute the MGFD:

- $\alpha = 1, \gamma = 1$ and $\sigma = 0$;
- $\alpha = 0.3, \gamma = 0.9$ and $\sigma = 30$;
- $\alpha = 0.4, \gamma = 1.2$ and $\sigma = 30$.

This resulted in three sets of features: PM_0, PM_1 and PM_2, respectively.

The converted speech detectors using PM_0, PM_1 and PM_2as features were trained and tested with the already described subsets of available data. The performance of the detectors is summarized in Tables 5.6, 5.7 and 5.8.

Log likelihood ratio	Test data (Acc. %)		
	Natural	GMM	US
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_GMM})$	73.1	77.8	49.8
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_US})$	70.2	43.8	64.0
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_mix})$	62.7	70.5	74.1

Table 5.6 Performance in accuracy rate of the GMM-based converted speech detector using PM_0 features in 9 test conditions

Log likelihood ratio	Test data (Acc. %)		
	Natural	GMM	US
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_GMM})$	81.8	82.6	10.2
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_US})$	54.3	38.9	73.7
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_mix})$	69.0	88.5	35.6

Table 5.7 Performance in accuracy rate of the GMM-based converted speech detector using PM_1 features in 9 test conditions

Log likelihood ratio	Test data (Acc. %)		
	Natural	GMM	US
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_GMM})$	83.9	83.8	29.1
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_US})$	73.9	51.5	67.7
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_mix})$	76.3	88.3	53.6

Table 5.8 Performance in accuracy rate of the GMM-based converted speech detector using PM_2 features in 9 test conditions

In the same way as the detector using the MM, this detector also shows a slightly poorer performance than the detector using MFCCs, confirming the difficulty to extract information related to long-term artifacts.

In the particular case of the performance of the detector using the PM_1, there was an extremely poor accuracy in detecting unit selection-based converted speech using the GMM-based converted model.

From the set of three features, the best performing one, on average and excluding the crossed detection case of unit-selection-based converted speech is the detector using PM_2 features.

It is relevant to mention that the parameter configuration for Eq. 40 used to extract the PM_2 is the same that it was used to extract the $MGDCCs_2$, which means that this may be the most adequate configuration for anti-spoofing mechanisms using phase spectral features.

5.4 Proposed converted speech detectors

The converted detection task studied throughout this chapter is, in its essence, a binary task, where the possible outcomes are either a prediction of natural speech or converted speech. As such, studying the performance of a discriminative modeling technique to address this task seems like a straightforward option. However that has not been the case as there are no reports in the literature for such techniques being used for converted speech detection. That may be the case because of the novelty of the topic and of anti-spoofing in speaker recognition tasks in general.

The purpose of this section is to overcome that knowledge gap by introducing the use of SVMs as the discriminative model approach to use for converted speech detection. A motivation for using SVM instead of another discriminative technique comes from its popularity as is a standard by which other methods are usually compared. Additionally, SVM-based methods have been outperforming log likelihood ratio-based methods in SV problems [23].

The goal of an SVM is to, given a training set of labeled, two-class examples, estimate a hyperplane that maximizes the separation of the two classes, after projecting them to a high dimensional space via Kernel function. After that, an SVM is used to predict the class of an unknown example by projecting the example into the same space as the hyperplane and determining in which side of the hyperplane it falls on. A more detailed description of the SVM algorithms and model can be found in Appendix A.

In order to evaluate the benefits of using SVMs for converted speech detection, several experiments are performed in this section. SVM models are estimated, using natural data as examples of the positive class and converted data (GMM-based, unit selection-based or mixed, exactly as in the detectors described in section 5.3) as examples of the negative class. The features used by the models are meant to ease the detection of spectral artifacts: short-term in the magnitude spectrum; short-term in the phase spectrum; long-term in the magnitude spectrum and long-term in the phase spectrum.

To ensure comparability with the performances of the GMM-based converted speech detectors implemented in section 5.3.1, the same features, training and testing data were used in the experiments carried out.

5.4.1 Proposed compact feature representation

In speech processing related tasks, a typical speaker's speech file containing one or more utterances lasts at least a few seconds and up to several minutes. On the other hand, most of the feature extraction processes employ frame-level approaches, using speech windows with no more that tens of milliseconds. Even the long-term features explored in section 4.4.2 do not require more than a second of speech. As such, for a given speech file of a speaker, and a given feature extraction process, the usual outcome of the process applied to the whole file is a matrix of $N \times C$ coefficients, where *N* is the number of frames in the file and *C* is the number of coefficients of the feature vectors.

As a very simple example, a possible feature matrix size for a speech file can be estimated by assuming a one minute file with no silence and 12 MFCCs extracted every 10 ms with 0% overlap from the previous frame as features. For this case the feature matrix would have $6,000 \times 12$ entries.

This feature representation ensures a detailed characterization of the speech signal for each moment in time, which is an important concern to have in many speech processing related tasks, for example in phoneme recognition. It also allows for a time analysis of the feature evolution, which may be important in tasks such as speech-to-text. However, for the particular task of converted speech detection, there is no need for such detailed characterization of speech in each moment, as the artifacts in the converted speech are present throughout the whole utterances and speech files. The premise for assuming so is that features extracted from converted speech are globally shifted towards a new distribution that is characteristic of the converted speech, compared to those of natural speech. These characteristic distributions may be a repercussion of the systematic occurrence of converted speech artifacts.

This section proposes the use of a lighter feature representation as an alternative to the "full" matrix representation used in nearly all the speech related applications. This representation can be applied to any features. Deriving the compact feature representation can be achieved, for a given speech file and a given feature extraction process, as follows:

1) Perform feature extraction as usual and obtain an $N \times C$ matrix, where N is the number of feature vectors and C is the number of coefficients in each feature vector;
2) Compute the mean for each coefficient over all the feature vectors and obtain a an $1 \times C$ means vector;
3) Repeat the previous step but instead of the mean, compute the standard deviation to obtain a an $1 \times C$ standard deviations vector;
4) Concatenate the means and standard deviations vectors to form a $1 \times 2C$ vector.

This $2C$-dimensional vector represents the whole speech file, and comparatively to the full representation, it implies a decrease the number of coefficients of approximately 10^3 times per minute of speech.

One of the advantages of such representation is that it reduces the training time of the model from several hours to a few seconds.

5.4.2 Experimental results

The same experiments as those described in section 5.3.1 were reproduced in this section with the only differences being the use of SVMs instead of GMMs as the modeling approach and the use of compact feature representation instead of the typical "full" representation for which results have already been reported in the literature.

The training and testing of the proposed detectors was made with the same subsets of available data used for training and testing the GMM-based converted speech detectors in section 3.2. A total of 9 conditions were contemplated, where each of the three test sets was tested against a model trained with natural and one of the three converted speech subsets.

In this study, after some development experiments, the algorithm chosen to perform optimization of the hyperplane and selection of support vectors was sequential minimal optimization (SMO) and the kernel chosen was the linear.

- *Detecting short-term magnitude spectrum artifacts:*

The proposed SVM-based converted speech detector implemented to detect short-term artifacts in the magnitude spectrum used MFCCs as features, following the extraction process described in section 2.2.1. The MFCCs were extracted in 20ms frames, updated every 10ms. Each feature vector had 12 MFCCs, no log-energy and no temporal derivatives. A VAD algorithm eliminated the feature vectors corresponding to silence. For each available speech file, feature representation conversion into compact representation, as described in section 5.4.1, was applied so that it only generated one 24-dimensional feature vector.

Three SVMs were trained using data from natural speakers as examples for the positive class, and data from either GMM-based, unit selection-based or mixed converted speakers as examples for the negative class. Each of the three trained SVMs was tested against natural data, GMM-based converted data and unit selection-based converted data.

The performance of the proposed SVMs is measured in accuracy rate and is summarized in Table 5.9. Each entry of the table corresponds to an experiment in similar conditions as the ones using GMM as modeling technique in Table 5.1.

Train data		**Test data (Acc. %)**		
Positive class	**Negative class**	**Natural**	**GMM**	**US**
Natural	*Convered (GMM)*	81.4	92.1	20.4
Natural	*Convered (US)*	67.4	83.9	81.8
Natural	*Convered (mix)*	70.5	92.3	70.1

Table 5.9 Performance in accuracy rate of the SVM-based converted speech detector using MFCCs in 9 test conditions

Comparing the results of this detector, using SVMs as the modeling technique and the compact feature representation of MFCCs, with the corresponding GMM-based detector, it is possible to observe that the overall performance is slightly poorer and that the issues found previously with crossed detection are still present.

- *Detecting short-term phase spectrum artifacts:*

Nine SVM models were trained, subdivided in three groups of three, where each group used a different feature, the GDCCs, $MGDCCs_1$ or $MGDCCs_2$. The features were extracted according to the process described in section 4.4.1, using 20ms frames and 10ms of overlap. Only the first 12 coefficients were kept. In addition, the parameter configuration of α, γ and σ was the same as in the GMM-based detectors:

- $\alpha = 1, \gamma = 1$ and $\sigma = 0$;
- $\alpha = 0.3, \gamma = 0.9$ and $\sigma = 30$;
- $\alpha = 0.4, \gamma = 1.2$ and $\sigma = 30$,

respectively. The features were compactly represented.

Each of the SVM models was tested against natural and converted data. The results for the GDCCs, $MGDCCs_1$ and $MGDCCs_2$ are shown in Tables 5.10, 5.11 and 5.12, respectively. The performance of these SVM-based converted speech detectors can be compared with the corresponding GMM-based ones in Tables 5.2, 5.3 and 5.4.

Train data		**Test data (Acc. %)**		
Positive class	**Negative class**	**Natural**	**GMM**	**US**
Natural	*Convered (GMM)*	97.3	97.2	71.1
Natural	*Convered (US)*	98.2	52.9	99.5
Natural	*Convered (mix)*	97.6	97.6	98.1

Table 5.10 Performance in accuracy rate of the SVM-based converted speech detector using GDCCs in 9 test conditions

Train data		**Test data (Acc. %)**		
Positive class	**Negative class**	**Natural**	**GMM**	**US**
Natural	*Convered (GMM)*	89.9	94.6	38.4
Natural	*Convered (US)*	78.0	96.0	85.9
Natural	*Convered (mix)*	83.1	96.5	74.3

Table 5.11 Performance in accuracy rate of the SVM-based converted speech detector using $MGDCCs_1$ in 9 test conditions

Train data		**Test data (Acc. %)**		
Positive class	**Negative class**	**Natural**	**GMM**	**US**
Natural	*Convered (GMM)*	97.9	96.4	30.7
Natural	*Convered (US)*	87.9	98.7	90.1
Natural	*Convered (mix)*	91.1	98.7	85.3

Table 5.12 Performance in accuracy rate of the SVM-based converted speech detector using $MGDCCs_2$ in 9 test conditions

Once again, the performances achieved using models trained with GDCCs, $MGDCCs_1$ and $MGDCCs_2$ features are better that the ones achieved with the MFCCs, which supports the hypothesis that the phase spectrum contains more (or at least more evident) artifacts than the magnitude spectrum given the fact that the phase spectrum is ignored in the VC process.

The SVM-based detectors outperformed the corresponding GMM-based ones in nearly every test condition. Moreover, the performance achieved by the SVM-based detector using the

GDCCs was the best one so far. Detection rates of over 95% show the usefulness of GDCCs for converted speech detection tasks.

- *Detecting long-term magnitude spectrum artifacts:*

Three more SVMs were trained in a similar fashion to what has been described. The features used in this set of experiments were the MM features and were extracted according to section 4.4.2. The experimental details were the same as described in section 5.3.1. The features were converted into a compact representation so that each speech file used resulted in a 20-dimensional feature vector.

The detectors were tested against the same subsets of data. The performance is summarized in Table 5.13 and is comparable to the one shown in Table 5.5, which is the equivalent experiment for GMM-based detectors.

Train data		**Test data (Acc. %)**		
Positive class	**Negative class**	**Natural**	**GMM**	**US**
Natural	*Convered (GMM)*	73.5	83.8	84.1
Natural	*Convered (US)*	84.0	64.3	90.6
Natural	*Convered (mix)*	76.7	78.6	91.3

Table 5.13 Performance in accuracy rate of the SVM-based converted speech detector using MM features in 9 test conditions

On average this SVM-based detector using MM features yields a better performance than the corresponding GMM-based one. Additionally, it is important to note that this detector achieved the best performance in the particular test condition of unit selection-based converted speech tested against a model trained with natural and GMM-based converted speech, which shows that this model and feature combination provide a robust alternative for converted speech detection.

- *Detecting long-term phase spectrum artifacts:*

The last experiments were performed in order to detect long-term artifacts in the phase spectrum with the proposed modeling approach and feature representation. The features used in this set of experiments were the PM features that were extracting following what is described in section 4.4.2. the PM feature were extracted three times each using a different parameter configuration for Eq. 40:

- $\alpha = 1, \gamma = 1$ and $\sigma = 0$;
- $\alpha = 0.3, \gamma = 0.9$ and $\sigma = 30$;
- $\alpha = 0.4, \gamma = 1.2$ and $\sigma = 30$.

The remaining experimental details were similar to the equivalent experiments with GMM-based detectors for long-term phase spectrum artifacts. The resulting features, PM_0, PM_1 and PM_2, respectively were represented in a compact way after converting them as in section 5.4.1.

The SVM-based converted speech detectors using PM_0, PM_1 and PM_2 as features were trained and tested with the usual subsets of available data. The performance of the detectors is summarized in Tables 5.14, 5.15 and 5.16. The performance of the equivalent GMM-based detectors is in Tables 5.6, 5.7 and 5.8.

Train data		Test data (Acc. %)		
Positive class	**Negative class**	**Natural**	**GMM**	**US**
Natural	*Convered (GMM)*	75.1	83.3	83.7
Natural	*Convered (US)*	82.9	61.3	91.0
Natural	*Convered (mix)*	75.7	78.2	91.1

Table 5.14 Performance in accuracy rate of the SVM-based converted speech detector using PM_0 features in 9 test conditions

Train data		Test data (Acc. %)		
Positive class	**Negative class**	**Natural**	**GMM**	**US**
Natural	*Convered (GMM)*	78.4	77.9	16.7
Natural	*Convered (US)*	52.9	44.3	83.6
Natural	*Convered (mix)*	52.7	72.4	79.6

Table 5.15 Performance in accuracy rate of the SVM-based converted speech detector using PM_1 features in 9 test conditions

Train data		Test data (Acc. %)		
Positive class	**Negative class**	**Natural**	**GMM**	**US**
Natural	*Convered (GMM)*	81.2	83.4	24.0
Natural	*Convered (US)*	67.2	39.5	72.6
Natural	*Convered (mix)*	69.2	84.3	59.8

Table 5.16 Performance in accuracy rate of the SVM-based converted speech detector using PM_2 features in 9 test conditions

Comparing the performance of the SVM-based detectors with the corresponding GMM-based detectors, one can see that the one using the PM_0 outperformed the corresponding one and the remaining yielded similar performances.

Comparing the performance of these three detectors among themselves, one can observe that the one using PM_0 features also outperformed the remaining two on most of the test conditions. Moreover, it provided reasonable detection accuracies for the test conditions concerning crossed method detection.

The model and feature combination of SVM and PM_0 is another alternative for robust converted speech detection.

5.5 Fusion of converted speech detectors

In speaker recognition related tasks and also in other detection and recognition problems in other fields, it is generally accepted that by fusing several standalone systems that perform the same task, the overall system performance will probably improve. It is based in this premise and in the suggestion made by [7] that the short- and long-term features carry complementary information that the possibility of score fusion of different detectors is introduced in this section.

The fusions described in sections 5.5.1 and 5.5.2 were achieved using the fusion algorithms implemented in the BOSARIS toolkit for MATLAB [67], which performs linear logistic regression (LLR) to fuse multiple sub-systems of binary classification. The fusion to produce the fused score, l, follows:

$$l = \sum_i \alpha_i s_i + b, \tag{41}$$

where α_i is the weight for the sub-system i and b is the offset. The parameters were trained in a development data set using a sort of 2-fold cross-validation [68]: development data is randomly split in two halves, one for parameter estimation and the other for assessment. This process is repeated using 10 different random partitions and the mean of the systems' performance can be computed. For the final submission, no partition of the data was made and all the development data was used to simultaneously calibrate the LLR fusion parameters.

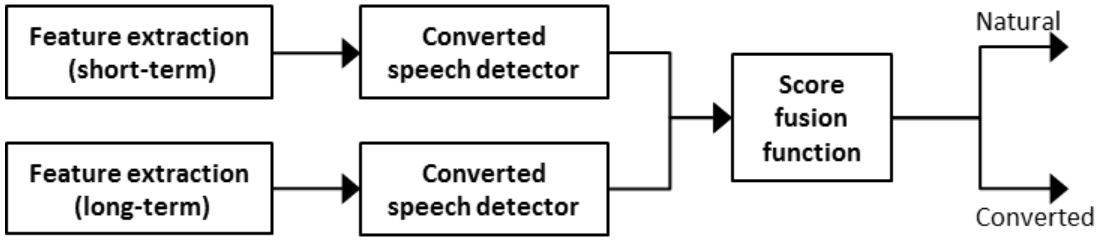

Fig. 5.1 Proposed fusion process for converted the speech detectors

The total of possible combinations of two systems of short and long term-features is 16, given the number of features already studied (4 short-term and 4 long-term). However, only the most relevant fusions (the ones yielding the best results) will be covered in a chapter so most of them will be overlooked. It should be mentioned that over 120 different converted speech detector combinations were considered during this study.

In the following sections only selected detectors using the pairs of features with the best performances will be addressed. The score fusions of the sub-systems using the following features were contemplated:

1) GDCCs and MM features
2) GDCCs and PM_0 features

5.5.1 Fusion of state-of-the-art converted speech detectors

The first featured experiment on score fusion of multiple sub-systems uses the GMM-based converted speech detector using GDCCs as features and the GMM-based detector using MM features.

The performances of the fused systems for the 9 test conditions that were previously studied are presented in Table 5.17. The scores are measured in accuracy rate, as usual.

Log likelihood ratio	Test data (Acc. %)		
	Natural	GMM	US
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_GMM})$	94.5	96.3	78.0
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_US})$	94.8	97.8	89.6
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_mix})$	96.7	96.0	94.6

Table 5.17 Performance in accuracy rate of the fused GMM-based converted speech detectors using MM features and GDCCs in 9 test conditions

From Table 5.17, it can be seen that there is an improvement of the accuracy detection for all the test conditions, comparing both with the GMM-based converted speech detector using the GDCCs as well as the one using the MM features (Tables 5.2 and 5.5). The improvement is particularly significant for the detector using the MM features.

The improvements achieved are a proof of the complementary information contained in short- and long-term features.

Another relevant observation is that the crossed method detection is achieving better performances that seen in standalone detectors.

The next sub-system fusions is accomplished using the GMM-based converted speech detector using GDCCs features and the GMM-based converted speech detector using the PM_0 features.

The performances for the 9 test conditions are presented in Table 5.18.

Log likelihood ratio	Test data (Acc. %)		
	Natural	GMM	US
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_GMM})$	94.6	96.6	78.7
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_US})$	94.6	97.6	89.4
$\log p(X\|\lambda_{natural}) - \log p(X\|\lambda_{converted\_mix})$	96.3	95.8	95.1

Table 5.18 Performance in accuracy rate of the fused GMM-based converted speech detectors using PM_0 features and GDCCs in 9 test conditions

As in the previous sub-systems fusion, this one also achieved an improved performance than any of the two standalone sub-systems used. The overall performance is very similar, hence it is unclear which of the two options is better.

5.5.2 Fusion of proposed converted speech detectors

Traditionally, the output of an SVM for inputted test data is a predicted label, for example 0 or 1. This label is assigned depending on which side of the separating hyperplane do the test input features fall on. In practice, the predicted label corresponds to the sign of $f(x)$ in Eq. 43. the term $f(x)$ itself is the distance of test features to the hyperplane, which is a more detailed

outcome. Hence, in the following experiments, the output of the SVMs will be considered to be the distance $f(x)$ to the hyperplane in order to allow a finer score fusion of the two sub-systems.

The first score fusion of the sub-systems using the proposed detectors was accomplished by using the SVM-based converted speech detector using the GDCCs as features and the one using the MM features. In both cases the features were represented in a compact way. The 9 test conditions were evaluated and the performance of the fused systems can be found in Table 5.19. The performance of the equivalent state-of-the-art fused systems can be found in Table 5.17.

Train data		**Test data (Acc. %)**		
Positive class	***Negative class***	**Natural**	**GMM**	**US**
Natural	*Convered (GMM)*	98.4	97.7	92.8
Natural	*Convered (US)*	85.2	84.4	98.4
Natural	*Convered (mix)*	98.2	97.8	98.2

Table 5.19 Performance in accuracy rate of the fused SVM-based converted speech detectors using MM features and GDCCs in 9 test conditions

The fusion of the sub-systems resulted in an improvement in the performance, comparing to the standalone detectors (Tables 5.10 and 5.13). The accuracy rates for this proposed fused converted speech detector are the highest of all the experiments, so far. Crossed method detection achieved accuracies comparable to same method detection or natural detection.

Next, the fusion of another pair of proposed detectors was made. The detectors used were the SVM-based converted speech detector using GDCCs as features and the one using PM_0. The performance of new fused system for the 9 test conditions are shown in Table 5.20. The corresponding performance table for the GMM-based fused system is Table 5.18.

Train data		**Test data (Acc. %)**		
Positive class	**Negative class**	**Natural**	**GMM**	**US**
Natural	*Convered (GMM)*	98.3	97.8	93.1
Natural	*Convered (US)*	84.1	84.6	98.4
Natural	*Convered (mix)*	98.1	97.2	98.0

Table 5.20 Performance in accuracy rate of the fused SVM-based converted speech detectors using PM_0 features and GDCCs in 9 test conditions

Once again, the fusion of the two sub-systems resulted in an improvement of the accuracy detection. This proposed fused converted speech detector outperformed the corresponding GMM-based one. In fact, the performance of this fused detector is very similar to the performance of the previous detector which was also observed in the two fused GMM-based detectors. In this detector crossed method detection is also successful.

It remains unclear which long term feature, MM features of PM_0 features, are better for the implemented converted speech detectors.

5.6 Results discusison

Some preliminary discussion about the experiments carried out in sections 5.35 to 5.5. has already been done as experimental details were being described and the results were being presented. The goal of this section is to provide a broader discussion about the achievements throughout the chapter. In particular, the following topics are important to address:

- *Influence of the reshaping parameters of the MGDF:*

The MGDF is a modification of the GDF as explained in section 4.4.1. In order to achieve this modification, three parameters are introduced in the GFD to reshape it, as seen in Eq. 40. These parameters are used to accentuate the fine structure of the GFD and to reduce its spiky nature. Several studies have reported improved results by applying modifications to the GDF [7] [66].

Throughout this study, several parameter combinations were studied, some are not even mentioned in this document. Such has helped to gain some insight on the influence of each parameter and how to tune it. However, in disagreement with the literature, the best performances for the experiments performed in this study were achieved without applying the modification to the GDF.

- *State-of-the-art vs. proposed converted speech detectors:*

The state-of-the-art converted speech detectors that were implemented in section 5.3 follow the works of [8], where short- and long term- features extracted from the magnitude and MGDF spectrum were used to build GMM-based models of natural and converted speech, whereas the proposed converted speech detectors used SVMs as the modeling technique.

From results presented throughout sections 5.3 and 5.4 are summarized in Fig. 5.2. One can see that the proposed converted speech detectors outperform the state-of-the-art ones in most cases. One notable exception occurred with the detectors trained with MFCCs, where the GMM-based detectors were more successful. Furthermore, the proposed converted speech detectors can in some cases successfully address the cross detection issue, which does not happen with any GMM-based converted speech detector.

The standalone detector that showed the best performance was one of the proposed SVM-based ones, particularly the one using GDCCs as features and trained with natural and mixed converted data, which achieved an accuracy of 97.8%.

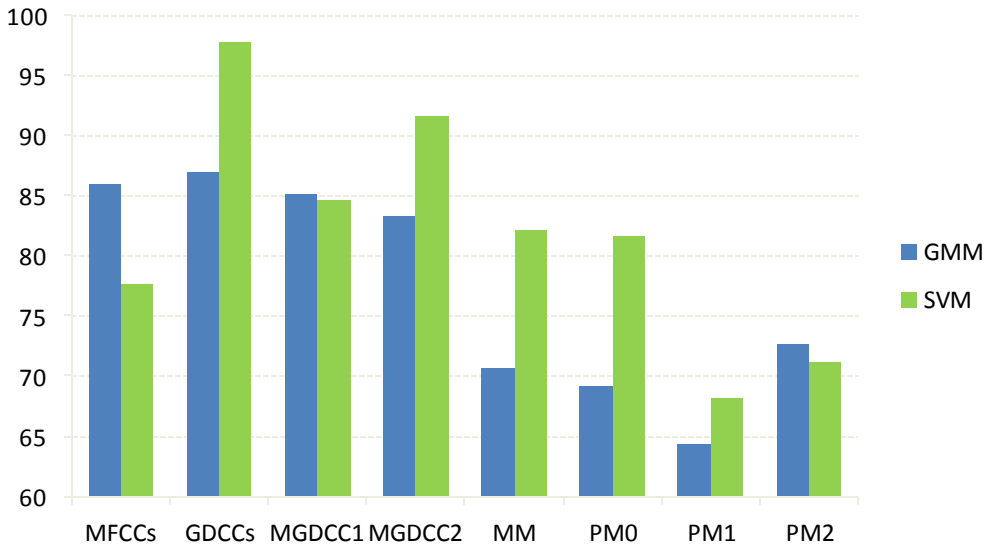

Fig. 5.2 Comparison of the performance in accuracy rate of GMM-based and SVM-based standalone converted speech detector for each feature assuming the training condition where mixed converted data is available

Regarding the performance of the fused converted speech detectors, which is summarized in Fig. 5.3, it's clear that the fusions achieved with the proposed SVM-based converted speech detectors yielded better performances. The best fused converted speech detector was SVM-based, featuring the MM features as the long-term feature and the GDCCs and the short term-features and trained with natural a mixed converted data. This detector achieved and accuracy of 98.1%, which is 2.3% more than the best fused GMM-based converted speech detector.

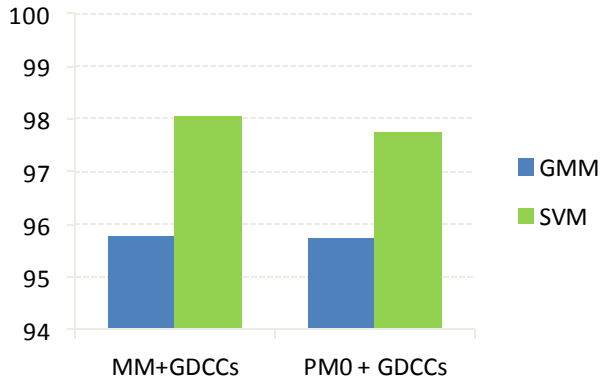

Fig. 5.3 Comparison of the performance in accuracy rate of GMM-based and SVM-based fused converted speech detectors for each feature pair assuming the training condition where mixed converted data is available

- *Standalone vs. fused converted speech detectors:*

The fusion of several sub-systems in automatic verification and detection tasks in a common practice and usually brings improvements to the performances previously achieved by the standalone systems. In section 5.5 some experiments regarding the fusion of the already described detectors were made. A summary of the performances can be found in Fig. 5.4.

Concerning the GMM-based converted speech detectors, the fusions of sub-systems caused a significant improvement in the performance of the two detectors used. The best GMM-based fused detector was achieved fusing the ones based on MM features as the long-term features and GDCCs as the short-term features and the training included data from natural speakers and mixed converted speakers. The performance of this detector was of 95.8% in accuracy. It is 25.1% more accurate than the corresponding sub-system using MM features and 8.8% more accurate than the one using GDCCs.

Regarding the fusion of the SVM-based converted speech detectors, it can be seen that it did not bring any significant improvement comparatively to the best performing sub-system: in the best of the two cases (the fused SVM-based detector used MM features as the long term features and GDCCs as the short term features and the training data included natural and mixed converted data) the improvement relatively to the sub-system using GDCCs was only from 98.0% to 98.1%.

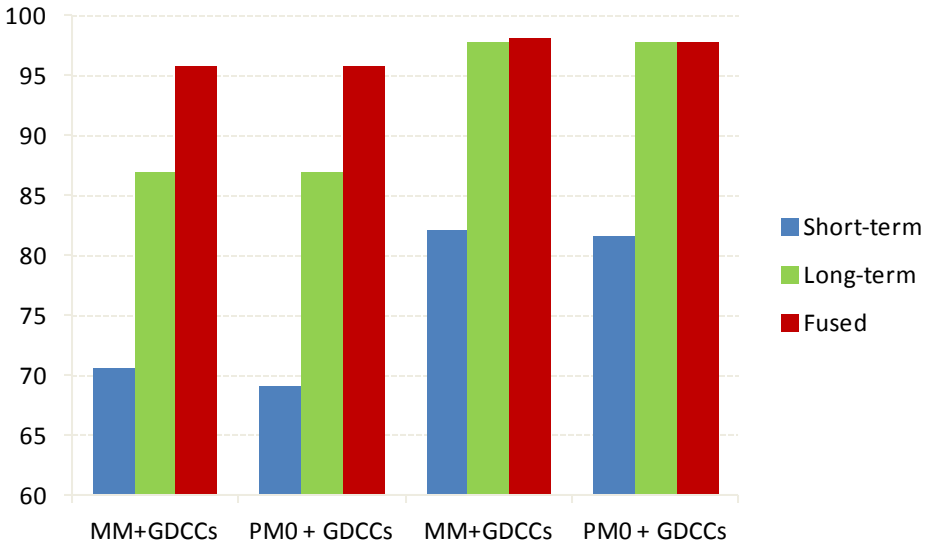

Fig. 5.4 Comparison of the performance in accuracy rate of the fused converted speech detectors and the corresponding sub-detectors for each feature pair and assuming the training condition where mixed converted data is available

- *Best performing converted speech detector vs. other reported experiments:*

As it was already mentioned, the best performing converted speech detector built in this study was the fused SVM-based detector resultant from the fusion of the sub-systems using the MM features and the GDCCs. Assuming the training conditions where natural and mixed converted speech were available, this system averaged 98.1% detection accuracy.

Other converted speech detectors have already been proposed in other studies. For example, in [9], De Leon proposed a detector that achieved an accuracy rate of 88%. In [8], based on the work of De Leon et al., Wu et al. proposed another detector which claimed to have 99% accuracy. The most recent study on converted speech detectors at the time of writing is [10], where Alegre et al. used a detector which achieved a performance of 97% accuracy.

The performances of these detectors reported in the literature are not directly comparable, as there are many differences in several important variables, such as the type of training data available or the difficulty of the trials set. As such, it would not be correct to determine which is the best detector solely by comparing the reported results. However, this study reproduced some of the state-of-the-art converted speech detectors that showed the best performances already reported and the proposed detectors outperform the state-of-the-art ones for the same task. As such, there is a strong possibility that the proposed converted speech detectors are among the most successful reported so far.

- *Originalities introduced:*

Previously to this study, all of the converted speech detectors reported in the literature used the *de facto* standard GMM as the modeling technique. This is the first study where an alternative modeling approach is used. The choice of the proposed modeling technique, the SVMs, was made with the particular converted detection task in mind, and the results shown an improved performance with the proposed approach.

Additionally, the compact feature representation is also a novelty, given that there are no reports of performing such transformations to the features. This makes the models of this proposed detector one of the simplest (in terms of amount of data needed to estimate the model) yet accurate ever estimated.

6 Speaker verification system with anti-spoofing mechanisms

6.1 Introduction

Anti-spoofing for SV systems is a fairly new topic in speech processing community, particularly concerning attacks from converted speech (as opposed to human mimicry, for example). Nevertheless there are already some publications that describe successful experiments and approaches that address this problem. Most of them focus on introducing an anti-spoofing mechanism as a post-processing module in an SV system in order to decrease the high FAR caused by the spoofing attacks. This option has added flexibility as the module is totally integrable with any existing SV system.

This brief chapter proposes an SV system including anti-spoofing mechanisms, following the state-of the-art tendency of adding a post-processing module to the SV system. The SV system that was used was the same one that served as a baseline in chapter 3. The anti-spoofing mechanist was based on the converted speech detectors evaluated in chapter 5. The chosen converted speech detector was the one that showed the best performance, hence the one that resulted from the fusion of the proposed SVM-based detectors using MM features and GDCCs.

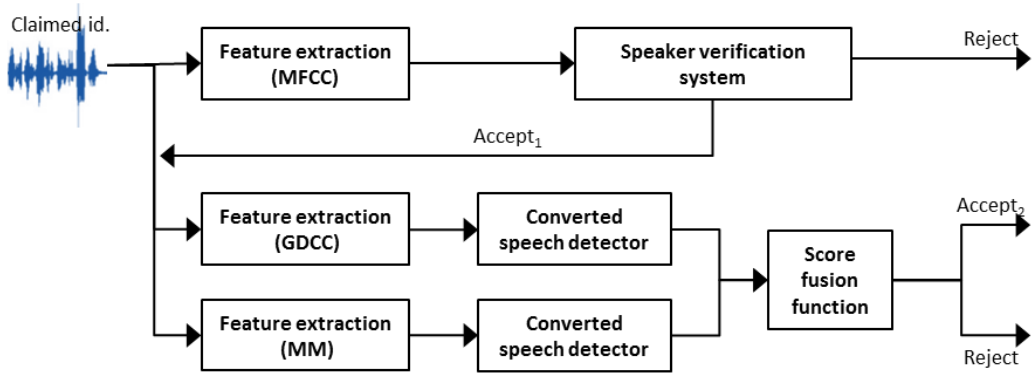

Fig. 6.1 SV system with the proposed anti-spoofing mechanism based on a fusion of two converted speech detectors

In this proposed SV system with anti-spoofing mechanisms, an unknown utterance is fed to the SV system, which treats it normally: extracts the features characterizing the utterance, uses them to compute the corresponding i-vectors and finally, makes the decision of accepting or rejecting the utterance as belonging to the target speaker by cosine similarity. If the SV system rejects the utterance the verification trial ends. However, if the utterance is accepted by the verification system, that decision will not be considered final, as it could be a false acceptance from a converted speech trial. The utterance is reused for a new process of feature extraction. At this point both MM features and GDCCs are extracted, each one to be fed to the corresponding SVM-based converted speech detector. The representation of the features is converted into the compact one, as the proposed SVM-based detectors require. The output scores of the two converted speech detectors are then combined using a pre-trained fusion

function. The output of the fusion function is compared to a decision threshold (which is set to zero) and the utterance is classified as natural if the fused score is above the threshold, otherwise it is classified as converted. If the SV system accepted the utterance and the anti-spoofing mechanism determined it belonged to a natural speaker, the utterance is accepted. On the other hand, if the SV system accepted the utterance but the anti-spoofing mechanist determined that the utterance belonged to a converted speaker, the utterance is rejected (considered a converted impostor trial). In practice this corresponds to manipulating the score given by the SV systems to each trial as follows:

- For each utterance rejected by the SV system, do nothing
- For each utterance accepted by the SV system and classified as natural by the anti-spoofing mechanism, do nothing
- For each utterance accepted by the SV system and classified as converted by the anti-spoofing mechanism, re-assign a new score below the decision threshold of the SV system so that it is rejected.

6.2 Corpora

The experiments carried in this chapter used the same corpora as previously mentioned in section 3.2, consisting of two natural speech corpora, one converted speech corpus achieved using GMM-based VC methods and one converted speech corpus achieved using unit selection-based VC methods. The natural data was already slip in training and testing data and the converted data was randomly split into two sub-corpora one for training and another for testing.

6.3 Experimental results

The SV system previously used for evaluating the vulnerability of state-of-the-art SV systems against converted speech spoofing attacks was reused in this experiment, as already mentioned. As a reminder, the system was based on the i-vector technology and trained with data from the 1conv4w train condition of the NIST SRE2006 speech corpus, as described in section 3.2. All of this data is from natural speakers. The post-processing module for anti-spoofing was included in the system by being inputted all the utterances accepted by the SV system. The best performing converted speech detector was trained with natural data and mixed converted data, so the same conditions are adopted for the following experiments.

This SV system with anti-spoofing mechanisms was tested against the same three corpora of natural, GMM-based and unit selection-based converted speech that were described in section 3.2, organized in the same trial sets as the experiments carried out in section 3.3. Hence, the results from the experiments carried out in this chapter are comparable to those of chapter 3, particularly from Table 3.4. To simplify the comparison of the performances the results from Table 3.4 were included in Table 6.1.

The proposed anti-spoofing mechanism was supposed to mitigate the effects of the converted speech spoofing attacks. In order to understand how well the proposed anti-spoofing

mechanism performed, it was included a simulation of an ideal anti-spoofing mechanism in the experiments descried in this chapter. Given that the true identity and the naturalness (the characteristic of being natural or converted) of the speaker from every trial is information that is available, it is possible to simulate an ideal anti-spoofing mechanism by assigning the desired score to each trial.

As mentioned, nothing is done for a trail rejected by the SV system. The performance of the SV system with the proposed anti-spoofing mechanism and simulation of the performance of an ideal anti-spoofing mechanism tested against the baseline set of trials, a set of trials containing natural and GMM-based converted speech and a set of natural and unit selection-based converted speech are summarized in Table 6.1. The DET curves for the SV system performance with the proposed ant-spoofing mechanism and the simulated ideal one can be found in Fig. 6.2 and 6.3.

SV system	Test data	Miss %	False acceptance %	Converted trials misclassifications %
Without anti-spoofing mechanism	*Baseline (no converted speech)*	9.4	9.4	-
	Natural and GMM-based converted speech	9.4	33.7	56.6
	Natural and unit selection-based converted speech	9.4	38.5	65.6
With proposed anti-spoofing mechanism	*Baseline (no converted speech)*	11.0	9.4	-
	Natural and GMM-based converted speech	11.0	5.9	2.2
	Natural and unit selection-based converted speech	11.0	5.7	1.8
With ideal anti-spoofing mechanism	*Baseline (no converted speech)*	9.4	9.4	-
	Natural and GMM-based converted speech	9.4	4.8	0.0
	Natural and unit selection-based converted speech	9.4	4.8	0.0

Table 6.1 Performance of the SV system without with proposed and ideal anti-spoofing mechanism in miss rate, FAR and converted trials misclassification rate against natural and converted speech

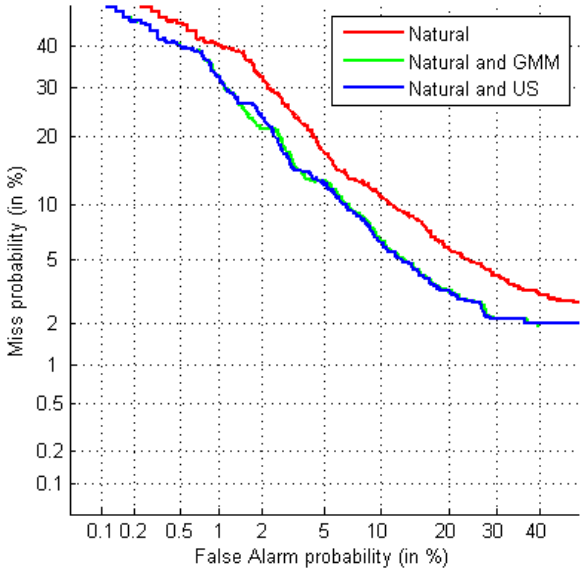

Fig. 6.2 DET curve of the performance of the SV system with the proposed anti-spoofing mechanism tested against natural and converted speech

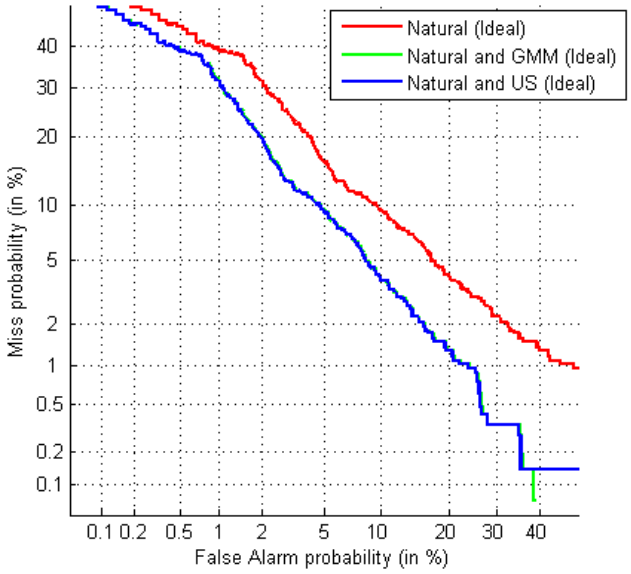

Fig. 6.3 DET curve of the performance of the SV system with an ideal anti-spoofing mechanism tested against natural and converted speech

The performance of the SV system with the proposed anti-spoofing mechanism shows a miss rate of 11%. Comparing to the miss rate of the baseline results there was an increase of 1.6%. This increased number of misses is a result of the misclassifications of natural speech trials by the anti-spoofing mechanism. On the other hand, the FAR remained constant for the baseline trials, which was something to expect given that the anti-spoofing mechanism, does not interfere with the utterances rejected by the SV system. The miss percentage remained at 11.0% for the remaining two test conditions and the FAR decreased from 33.7% to 5,9% and from 38.5% to 5.7% for the case of the trial set with GMM-based and unit selection-based converted speech, respectively. The DET curves in Fig. 6.2 show how similar the performance of the system for this two test conditions is. Relatively to converted misclassifications, there

was a significant decrease in it, from 56.6% to 2.2% in the case of GMM-based converted speech and from 65.6% down to 1.8% in the case of unit selection-based converted speech.

The performance of an ideal anti-spoofing mechanism for these tasks would yield a miss rate of 9.4% for every test condition, which means that no natural utterance would be misclassified by the anti-spoofing mechanism. Moreover, the FAR would be 9.4% for the baseline condition and 4.8% for both the test conditions containing converted speech. The misclassifications of the converted speech utterances would also 0.0%, as by definition.

Comparing the performance of the SV system with the proposed anti-spoofing mechanism with the simulated ideal anti-spoofing mechanism, it is possible to observe that the miss rate for natural trials is of 1.6% higher, while the FAR is 1.1% or less higher than in an ideal situation.

SV system	Test data	Miss %	False acceptance %	Converted trials misclassifications %
Without anti-spoofing mechanism	*Baseline (no converted speech)*	8.9	8.9	-
	Natural and GMM-based converted speech	8.9	33.0	52.9
	Natural and unit selection-based converted speech	8.9	38.3	60.7
With proposed anti-spoofing mechanism	*Baseline (no converted speech)*	10.7	8.9	-
	Natural and GMM-based converted speech	10.7	5.2	2.0
	Natural and unit selection-based converted speech	10.7	4.8	1.6
With ideal anti-spoofing mechanism	*Baseline (no converted speech)*	8.9	8.9	-
	Natural and GMM-based converted speech	8.9	4.1	0.0
	Natural and unit selection-based converted speech	8.9	3.9	0.0

Table 6.2 Performance of the female SV system without, with proposed and ideal anti-spoofing mechanism in miss rate, FAR and converted trials misclassification rate against natural and converted speech

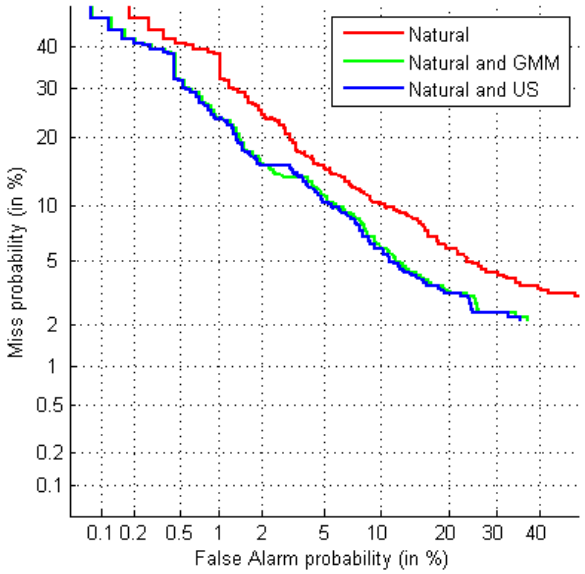

Fig. 6.4 DET curve of the performance of the female SV system with the proposed anti-spoofing mechanism tested against natural and converted speech

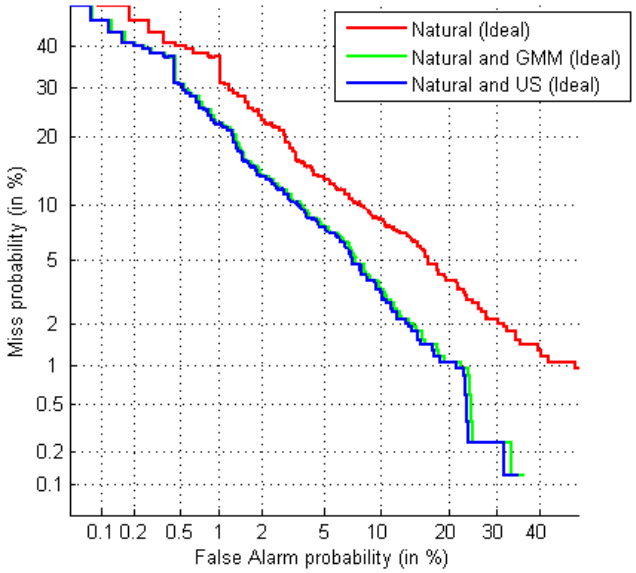

Fig. 6.5 DET curve of the performance of the female SV system with an ideal anti-spoofing mechanism tested against natural and converted speech

In Chapter 3 the results of the experiments carried out to evaluate the vulnerability of an SV system against a converted speech spoofing attack were presented separately for each gender. Then, in order to lighten the descriptions and reduce the amount of tables and plots in this document, the results from thereon were presented as a pooling of the female and male results for the female and male models, respectively. Now, upon reaching the final experiments carried out in this study, it is relevant to adopt the initial gender-dependent style of presenting the results in order to understand the differences and similarities between the performances of the system with anti-spoofing mechanisms for each gender. As such, Table 6.2 shows the performance of the SV system with the proposed anti-spoofing mechanism correspondent to the models (for both the SV system and the converted speech detectors) trained with females data and tested with female data. Table 6.2 also features a simulation of

an ideal anti-spoofing mechanism for the female trials. The equivalent results for male models and trials are presented in Table 6.4.

Fig. 6.3 shows the DET curve obtained for the performance of the female SV system with the proposed anti-spoofing mechanism tested against natural, GMM-based and unit selection-based converted speech. Fig. 6.4 shows the DET curve for the simulated performance of the female SV system with an ideal anti-spoofing mechanism. The equivalent DET curves for the performance of male systems can be found in Fig. 6.5 and Fig. 6.6.

Regarding the results for female models and comparing them with the baseline results for the female SV system without anti-spoofing mechanisms, it can be seen that the miss rate for all three trials sets is 10.7%, which is 1.8% more that the miss% of the same SV system without the anti-spoofing mechanism. The FAR remained constant for the baseline set of trials at 8.9% and decreased from 33.0% to 5.2% and from 38.3% to 4.8% for the sets of trials including GMM-based and unit-selection based converted speech, respectively. The misclassifications of converted trials were of 2.0% and 1.6% for GMM-based and unit selection-based converted speech trials, respectively.

SV system	Test data	Miss %	False acceptance %	Converted trials misclassifications %
Without anti-spoofing mechanism	*Baseline (no converted speech)*	9.9	9.9	-
	Natural and GMM-based converted speech	9.9	35.2	63.5
	Natural and unit selection-based converted speech	9.9	39.3	75.7
With proposed anti-spoofing mechanism	*Baseline (no converted speech)*	11.3	9.9	-
	Natural and GMM-based converted speech	11.3	6.8	2.4
	Natural and unit selection-based converted speech	11.3	6.8	2.0
With ideal anti-spoofing mechanism	*Baseline (no converted speech)*	9.9	9.9	-
	Natural and GMM-based converted speech	9.9	5.7	0.0
	Natural and unit selection-based converted speech	9.9	5.9	0.0

Table 6.3 Performance of the male SV system without, with proposed and ideal anti-spoofing mechanism in miss rate, FAR and converted trials misclassification rate against natural and converted speech

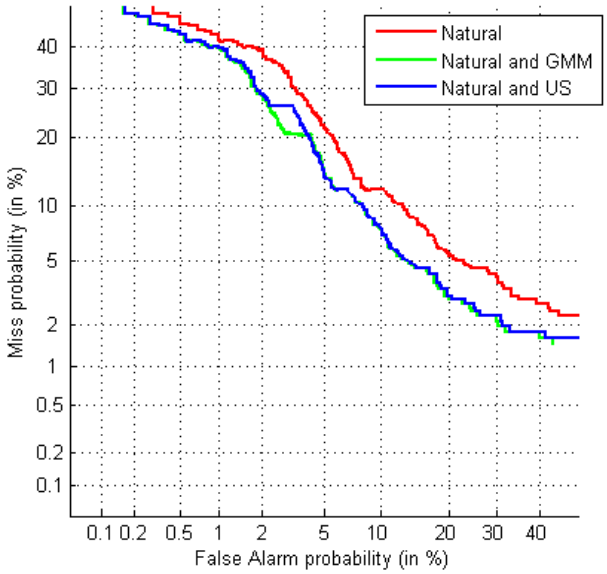

Fig. 6.6 DET curve of the performance of the male SV system with the proposed anti-spoofing mechanism tested against natural and converted speech

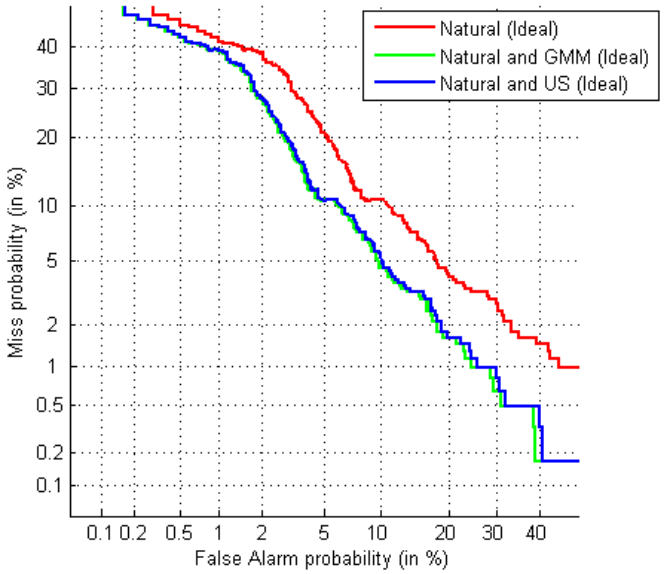

Fig. 6.7 DET curve of the performance of the male SV system with an ideal anti-spoofing mechanism tested against natural and converted speech

As for the results concerning the male models, the miss rate for the SV system with the proposed anti-spoofing mechanism was of 11.3%, an increase of 1.4% from the same SV system without the mechanism. The FAR decreased significantly, as seen in the female case, from 35.2% to 6.8% and from 39.3% to 6.8% for the set of trials with natural and GMM-based and unit selection-based converted speech, respectively. The miss classifications of converted trials also decreased, specifically from 63.5% to 2.4% and from 75.7% to 2.0% for GMM-based and unit selection-based converted speech trials, respectively.

Comparing the female and male performances, it is possible to observe that the performance of the SV system with the proposed anti-spoofing mechanism is slightly better in the female

case rather than the male, which comes in line with the results obtained in the experiments in section 3.3. The converted misclassifications are further reduced for the female experiments, where a score as low as 1.6% was achieved. Overall the results obtained for male and female are not too different between themselves as well as they are coherent with what had been seen in the experiments in section 3.3.

Overall, it can be seen that the proposed anti-spoofing mechanism proved to be an effective countermeasure to the converted speech spoofing attacks, leaving merely an average of 1.9% of all the trials remain misclassified. Achieving such results puts this anti-spoofing mechanism head to head with the best ones reported in the literature.

7 Conclusion

7.1 Discussion

The goal of this thesis was to contribute to the development of converted speech detectors as anti-spoofing mechanisms and enable them to be integrated into any SV system in order to increase its robustness against converted speech spoofing attacks.

The accomplishment of this goal encompassed several steps. The starting point was the implementation of a state-of-the art SV system and the evaluation of its performance against natural data from target and impostor speakers. The purpose of doing so was to establish a baseline performance for the SV system. This was followed by an evaluation of the performance of the same SV system when faced against two different converted speech corpora, based on GMM-based and unit selection-based VC methods. The comparison of the performance of the SV system in the baseline conditions and against converted speech data confirmed the vulnerability of the SV system.

This provided the motivation for the development of new anti-spoofing mechanisms. The most successful anti-spoofing mechanisms reported in the literature suggest that the best way to implement a countermeasure to spoofing attacks is to add a post-processing module to the SV system based on a converted speech detector. As such, several converted speech detectors already reported were reproduced in this thesis. The modeling technique chosen was the GMM model. The features used to characterize the speech were specifically chosen so that they capture the differences between natural and converted speech. These differences are mainly spectral artifacts introduced both in the magnitude spectrum and the phase spectrum of the converted speech during the VC process. These detectors were then tested against natural and converted data.

The best result for the GMM-based converted speech detector was of 87.0% average accuracy and corresponded to the detector trained with natural and mixed converted data, using GDCC features. This performance did not match the one reported by the authors in [7]. This may be because of further developments in the converted detectors that are not described in the report or because the detector was tested against an easier tasks.

The main contribution of this thesis was the development of new converted speech detectors based on a discriminative modeling technique, and on a novel compact feature representation, that allows much faster model training without any degradation of the performance of the detectors. These new detectors were trained with the same data and tested against the same trials as the GMM-based ones, in order to ensure comparability.

The best performing proposed converted speech detector was again the one using GDCC features and trained with natural and mixed converted data. However, the performance of this detector vastly outperformed the corresponding state-of-the-art one, achieving an average accuracy of 97.8%.

In order to make the performance of the detectors more robust, some experiments regarding the fusion of the developed converted speech detectors were also performed. The fusion of the detectors was achieved by pairing two detectors which used the same modeling technique and where one of them used a short-term feature and the other used a long term-feature. The goal of this fusion approach was to take advantage of the complementary information existing in short- and long-term features.

The best result of the fusion of the GMM-based detectors was achieved by training the detectors with natural and mixed converted data, and using both MM and GDCC features. The 95.8% accuracy represents a significant improvement over the performance of each of the sub-systems.

On the other hand, the SVM-based detectors did not show significant improvements after fusion. Nevertheless, the best fused converted speech detector was, once again, one of the proposed ones, particularly the one achieved by fusing the sub-systems trained with natural and mixed converted data and using both MM and GDCC features. This detector yielded a performance of 98.1% average accuracy.

After this extensive study on converted speech detectors, the final goal of creating an anti-spoofing mechanism was achieved by including the best converted speech detector as a post-processing module for the accepted utterances of the SV system. The performance of the system with the anti-spoofing mechanism was reevaluated, and compared to the simulated performance of an SV system with an ideal anti-spoofing mechanism. The proposed anti-spoofing mechanism was able to successfully mitigate the typical effects of a converted speech spoofing attack. The FAR decreased to acceptable values, specifically, to within 1.6% FAR of the ideal performance. Overall, only 1.9% of the trials accepted by the SV remained misclassified.

The following two publications resulted from this thesis:

- M. J. Correia, “Towards an Anti-spoofing System Based on Phase and Modulation Features,” in INESC-ID Tech. Reports, December 2013.
- M. J. Correia, A. Abad and I. Trancoso, “Preventing converted speech spoofing attacks in speaker verification,” in MIPRO 37th International Conference, Opatija, Croatia, May 2014.

7.2 Future work

The topic of anti-spoofing for SV is fairly recent, the first studies on imposture by synthetic speech dating back from only a decade ago. Hence, the amount of unexplored approaches and methods is far greater than in other areas. As such there are many possible future further developments that can come from this study. Only to name a few:

- Explore using detectors with new features, in order to increase robustness in adverse conditions;
- Experiment pairing more than two detectors, as it has been seen in the literature that usually, the more sub-systems are used in a fusion, the greater are the improvements in the overall performance;
- Experiment training the SV systems themselves with the features specifically chosen to detect spectral artifacts. An example would be to train a SV system with MFCCs and GDCCs, and evaluate if the SV system is less affected by the spoofing attack.

References

[1] J.P. Campbell Jr, "Speaker recognition: A tutorial," Proceedings of the IEEE, vol. 85, no. 9, pp. 437,1462, 1997.

[2] G. Fant, "Acoustic Theory of Speech Production," The Netherlands: Mouton-The Hague, 1970.

[3] E. Moulines, and Y. Sagisaka, "Voice Conversion: State of the Art and Perspectives," Speech Communication, vol. 16, no. 2, 1995.

[4] J.F. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," Interspeech, 2007.

[5] Q. Jin, A. Toth, A.W. Black, and T. Schultz, "Is voice transformation a threat to speaker identification?," Acoustics, Speech, and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on , vol. 1, no. 1, pp.4845,4848, 2008.

[6] Z. Wu, T. Kinnunen, E. Chng, H. Li, and E. Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case," Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific, vol. 1, no. 1, pp.1,5, 2012.

[7] Z. Wu, E. Chng, and H. Li, "Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition", Proceedings of Interspeech, 2012.

[8] Z. Wu, X. Xiao, E. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature", Acoustics, Speech, and Signal Processing (ICASSP), 2013 IEEE International Conference on , vol. 1, no. 1, pp.7234,7238, 2013.

[9] L. De Leon, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi, "Detection of synthetic speech for the problem of imposture," Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp.4844-4847, 2011

[10] F. Alegre, A. Amehraye, N. Evans, "Spoofing countermeasures to protect automatic speaker verification from voice conversion," Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on , vol.1 , no.1 , pp.3068,3072, 2013.

[11] J. Wolf "Efficient acoustic parameters for speaker recognition," Journal of The Acoustical Society of America, vol. 51, no. 2, pp.2044-2055, 1972.

[12] J. Ostrander, T. Hopmann, E. Delp, "Speech recognition using LPC analysis," Technical Report RSD-TR-1-82, University of Michigan, 1982.

[13] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," Journal of The Acoustical Society of America, vol. 87, no. 4, pp. 65-68, 1990.

[14] H. Hermansky, and N. Morgan, "RASTA processing of speech," Speech and Audio Processing, IEEE Transactions on , vol.2, no.4, pp.578,589, 1994.

[15] S. Furui, "Cepstral analysis technique for automatic speaker verification," Acoustics, Speech and Signal Processing, IEEE Transactions on , vol.29, no.2, pp.254,272, 1981

[16] K. Woo, T. Yang, K. Park, C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," Electronics Letters, vol. 36, no. 2, pp. 180–181, 2000.

[17] Q. Li, J. Zheng, A. Tsai, Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," Speech and Audio Processing, IEEE Transactions on , vol.10, no.3, pp.146,157, 2002.

[18] M. Marzinzik, B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," Speech and Audio Processing, IEEE Transactions on , vol.10, no.2, pp.109,118, 2002.

[19] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, K. Zechner, "Advances in automatic meeting record creation and access," Acoustics, Speech, and Signal Processing (ICASSP), 2001, IEEE International Conference on , vol.1, no., pp.597,600 vol.1, 2001.

[20] D. Reynolds, and R. Rose, “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, Speech and Audio Processing, IEEE Transactions on , vol.3, no.1, pp.72,83, 1995.
[21] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” Digital Signal Processing, vol. 10, no. 1, pp. 19,41, 2000.
[22] D. Reynolds, “An Overview of Automatic Speaker Recognition Technology”, Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on , vol.4, no.1, pp.IV-4072,IV-4075, 2002.
[23] W. Campbell, D. Sturim, and D. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” Signal Processing Letters, IEEE, vol.13, no.5, pp.308,311, 2006.
[24] R. Collobert, and S. Bengio, “SVMTorch: Support vector machines for large-scale regression problems,” Journal of Machine Learning Research, vol. 1, pp. 143–160, 2001.
[25] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Speaker and session variability in GMM-based speaker verification.” Audio, Speech, and Language Processing, IEEE Transactions on , vol.15, no.4, pp.1448,1460, 2007
[26] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, “Comparison of scoring methods used in speaker recognition with joint factor analysis,” Acoustics, Speech and Signal Processing (ICASSP), 2009. IEEE International Conference on , vol. 1, no.1 , pp.4057,4060, 2009.
[27] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, “Front-End Factor Analysis for Speaker Verification,” Audio, Speech, and Language Processing, IEEE Transactions on , vol.19, no.4, pp.788,798, 2011.
[28] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, “Score normalization for text-independent speaker verification systems,” Digital Signal Processing, vol. 10, pp. 42–54, 2000.
[29] A.F. Machado, and M. Queiroz, “Voice conversion: A critical survey,” Proceedings of Sound and Music Computing (SMC), 2010
[30] Y. Stylianou, “Voice transformation: a survey,” Acoustics, Speech and Signal Processing, (ICASSP)2009.IEEE International Conference on , vol.1 , no.1 , pp.3585,3588, 2009.
[31] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," Speech and Audio Processing, IEEE Transactions on , vol.6, no.2, pp.131,142, 1998.
[32] E. K. Kim, S. Lee, and Y. H. Oh, “Hidden Markov model-based voice conversion using dynamic characteristics of speaker,” Proceedings of EUROSPEECH, vol. 5, pp. 2519–2522, 1997.
[33] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, S. Narayanan, “Text-Independent Voice Conversion Based on Unit Selection,” Acoustics, Speech and Signal Processing (ICASSP), 2006. IEEE International Conference on , vol.1, no.1 , pp.I,I, 2006.
[34] M. Wilde, A. Martinez, “Probabilistic principal component analysis applied to voice conversion,” Signals, Systems and Computers, 2004. Conference Record of the Thirty-Eighth Asilomar Conference on , vol.2, no.1 , pp.2255,2259, 2004.
[35] S. Desai, E. Raghavendra, B. Yegnanarayana, A. Black, K. Prahallad, “Voice conversion using Artificial Neural Networks,” Acoustics, Speech and Signal Processing (ICASSP), 2009. IEEE International Conference on , vol.1 , no.1 , pp.3893,3896, 2009.
[36] M. Abe, S. Nakamura, K. Shikano, H. Kuwabara, “Voice conversion through vector quantization,” Acoustics, Speech, and Signal Processing (ICASSP), 1988. International Conference on , vol.1 , no.1 , pp.655,658 1988.
[37] D. Erro , A.Moreno, “Weighted Frequency Warping for Voice Conversion”, Proceedings of Interspeech , 2007.

[38] Y. Stylianou, O. Cappe, and E. Moulines, “Statistical methods for voice quality transformation,” Proceedings of the European Conference on Speech Processing and Technology, vol.1, no. 1, pp. 447,450, 1995.
[39] D. Sundermann , A. Bonafonte , H. Hoge, and H. Ney "Voice conversion using exclusively unaligned training data," Procesamiento del Lenguaje Natural, 2004.
[40] D. Erro, and A. Moreno, “Frame alignment method for cross-lingual voice conversion,” Proceedings of Interspeech, pp. 1969–1972, 2007.
[41] N. Campbell, and A. Black, “Prosody and the selection of source units for concatenative synthesis,” In J. van Santen, R. Sproat, J. Olive, and J. Hirschberg, editors, Progress in Speech Synthesis. Springer Verlag, 1995.
[42] A.J. Hunt, and A.W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” Acoustics, Speech, and Signal Processing (ICASSP), 1996. IEEE International Conference on , vol.1, no.1 , pp.373,376 vol. 1, 1996.
[43] F.Diego, A.Bonafonte, and A.Moreno, “Voice conversion of non-aligned data using unit selection,” TC-STAR Workshop on Speech to Speech Translation, 2006.
[44] T. Toda, A. W. Black, and K. Tokuda, “Acoustic-to-articulatory inversion mapping with Gaussian mixture model,” Proceedings of the International Conference on Spoken Language Processing, 2004.
[45] A. Toth, and A. Black, “Using articulatory position data in voice transformation,” in Workshop on Speech Synthesis, pp. 182–187, 2007.
[47] T. Matuko, K. Tokuda, K . Kobayashi, and S. Imai, “Imposture using synthetic speech against speaker verification based on spectrum and pitch,” in Proceedings of the International Conference on Spoken Language Processing, 2000
[48] P. Torres-Carrasquillo, E. Singer, M. Kohler, R. Greene, D. Reynolds, J. Deller Jr., “Approaches to language identification using Gaussian mixture models and shifted delta cepstral features” in Interspeech, 2002
[49] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of inter-speaker variability in speaker verification,” Audio, Speech, and Language Processing, IEEE Transactions on , vol.16, no.5, pp.980,988, 2008.
[50] I. Buhan, and P. Hartel, “The state of the art in abuse of biometrics”, 2005.
[51] Y. Lau, D. Tran, and M. Wagner, “Testing voice mimicry with the yoho speaker verification corpus,” in Knowledge-Based Intelligent Information and Engineering Systems. Springer, pp. 907-907, 2005.
[52] J. Villalba, and E. Lleida, “Preventing replay attacks on speaker verification systems,” in Security Technology (ICCST), 2011 IEEE International Carnahan Conference on. IEEE, 2011
[53] D. Matrouf, J.-F. Bonastre, and C. Fredouille, “Effect of speech transformation on impostor acceptance,” in Acoustics, Speech and Signal Processing (ICASSP), 2006. IEEE International Conference on, vol. 1, pp. I-I , 2006.
[54] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, “Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech,” in Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, pp. 4401,4404, 2012.
[55] L. Alsteris, and K.Paliwal. “Short-time phase spectrum in speech processing: A review and some experimental results,” Digital Signal Processing, vol. 1, no, 7, pp.578–616, 2007.
[56] H. Pobloth, and W. B. Kleijn, “On phase perception in speech,” Acoustics, Speech, and Signal Processing (ICASSP), 1999. IEEE International Conference on , vol.1, no.1 , pp.29,32 1999.
[57] G. Doddington, “Speaker recognition based on idiolectal differences between speakers,” Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech), 2001.

[58] L. Ferrer, H. Bratt, V. R. R. Gadde, S. Kajarekar, E. Shriberg, K. Sonmez, A. Stolcke, and A. Venkataraman, “Modeling duration patterns for speaker recognition,” Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech), 2003.
[59] G. Baudoin, and Y. Stylianou , “On the transformation of the speech spectrum for voice conversion,” Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP), vol. 3, no. 1, pp.1405,1408, 1996.
[60] K. Fujii, J. Okawa, and K. Suigetsu, “High individuality voice conversion based on concatenative speech synthesis,” World Academy of Science, Engineering and Technology, vol. 2, p. 1, 2007.
[61] A. Kain, and M. Macon, “Text-to-speech voice adaptation from sparse training data,” Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech), 1998
[62] J. Tribolet, “A new phase unwrapping algorithm,” Acoustics, Speech and Signal Processing, IEEE Transactions on , vol.25, no.2, pp.170-177, 1977
[63] B. Yegnanarayana, J. Sreekanth, A. Rangarajan, “Waveform estimation using group delay processing,” Acoustics, Speech and Signal Processing, IEEE Transactions on , vol.33, no.4, pp. 832- 836, 1985
[64] L. D. Alsteris, K. Paliwal. 2007, “Short-time phase spectrum in speech processing: A review and some experimental results,” Digit. Signal Process, vol. 17, no. 3 pp. 578-616, 2007
[65] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," IEEE Signal Processing Letters, vol. 6, no. 1, pp. 1-3, 1999.
[66] H.A. Murthy, and V. Gadde, “The modified group delay function and its application to phoneme recognition,” in Acoustics, Speech, and Signal Processing (ICASSP), 2003. IEEE International Conference, vol. 1, pp. I–68, 2003.
[67] “Bosaris toolkit [software package],” WWW page, March 2014, https://sites.google.com/site/bosaristoolki
[68] R. Fuentes “The BLZ Submission to the NIST 2011 LRE: Data Collection, System Development and Performance,” Proceedings of Interspeech, 2012.
[69] J. Platt, “Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines,” 1998

Appendix A – Support Vector Machines

SVMs are a set of supervised learning models and learning algorithms that are used as binary linear classifiers. The goal of an SVM is to, given a set of labeled, two-class data, find the separating hyperplane that maximized the distance to the two classes. The test examples will then be compared to the hyperplane and a predicted label will be associated to them according to which side of the hyperplane they fall on.

In practice that corresponds to, given a set of example-label pairs $(x_i, y_i), i = 1, \dots, l$ where $x_i \in \mathbb{R}^n$ and $y \in \{1, -1\}^l$, finding the solution of the optimization problem:

$$\min_{w,b,\xi} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{I}\xi_i \tag{42}$$

$$\text{Subject to } \; y_i(\boldsymbol{w}^T\phi(x_i) + b) \geq 1 - \xi_i, \;\; \xi_i > 0.$$

The training vectors x_i are mapped into a higher dimensional space by the function ϕ. An optimization algorithm such as SMO least squares or quadratic programming, finds a linear separating hyperplane with the maximal margin in this high dimensional space. $C > 0$ is the penalty parameter of the error term. Furthermore, $K(\boldsymbol{x}_i, \boldsymbol{x}_j) \equiv \phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}_j)$ is called the kernel function. The most used kernels are:

- Linear: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T\boldsymbol{x}_j$;
- Polynomial: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\gamma\boldsymbol{x}_i^T\boldsymbol{x}_j + r)^d, \gamma > 0$;
- Radial basis function (RBF): $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\gamma\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2\right), \gamma > 0$;
- Sigmoid, $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\gamma\boldsymbol{x}_i^T\boldsymbol{x}_j + r)$;

where γ, r and d are kernel parameters.

The trained SVM model is used to predict the label of a new example based on the function:

$$f(x) = \sum_{i=1}^{L}\alpha_i t_i K(x, v_i) + k \tag{43}$$

Where x is the input data; L is the number of support vectors, α_i and k are training parameters, v_i are the support vectors, obtained via an optimization process [67]. $K(\cdot,\cdot)$ is the kernel function, and t_i are the ideal outputs, with values ±1 depending on whether the accompanying support vectors belong to class 0 or 1. Overall, the parameters are subject to the constraint $\sum_{i=1}^{L}\alpha_i t_i = 0$.

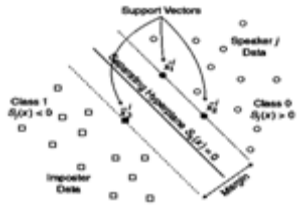

Fig. A.1 SVM separating hyperplane and its support vectors