ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2014-ENST-0082

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Signal et images »

présentée et soutenue publiquement par

Federico Leonardo ALEGRE

le 15 Décembre 2014

La protection des systèmes de reconnaissance de locuteur contre le leurrage

Directeur de thèse : **Nicholas W. EVANS**

Jury

M. John S. D. MASON, Professeur, Université de Swansea	Rapporteur
M. Driss MATROUF, Maître de Conférences HDR, LIA, Université d'Avignon	Rapporteur
M. Prénom NOM, Professeur, Institut EURECOM	Examinateur
M. Benoit FAUVE, Scientifique en chef, ValidSoft Ltd	Invité
M. Artur JANICKI, Maître de Conférences, Ecole Polytechnique de Varsovie	Invité

TELECOM ParisTech
école de l'Institut Télécom - membre de ParisTech

THÈSE

Spoofing & countermeasures for biometric speaker verification

Federico Leonardo ALEGRE

A doctoral dissertation submitted to:

TELECOM ParisTech

in partial fulfillment of the requirements for the degree of:

DOCTOR (PhD)

Specialty : Signal & Image

Jury :

Reviewers:
Prof. John S. D. MASON - Swansea University (UK)
Dr. Driss MATROUF - LIA - Université de Avignon (France)
Examiner:
Prof. Dirk SLOCK - EURECOM, Sophia Antipolis (France)
Invited guests:
Dr. Benoıt FAUVE - ValidSoft Ltd, London (UK)
Dr. Artur JANICKI - Warsaw University of Technology (Poland)
Supervisor:
Dr. Nicholas EVANS - EURECOM, Sophia Antipolis (France)

Abstract

As a result of the growing need for secure systems and services, the design of reliable personal recognition systems is becoming more and more important. In this context, biometric systems which use physiological and/or behavioural traits such as fingerprints, face, iris or voice for automatic recognition of individuals has a number of advantages over conventional authentication methods such as PINs, cards or passports.

In spite of the advantages, however, a growing body of independent work shows that all biometrics systems are vulnerable to subversion, either evasion in the case of surveillance or, as is the interest in this thesis, spoofing in the case of authentication. Surprisingly, there is only a small (but growing) body of work to develop countermeasures which can offer some protection from spoofing attacks.

This thesis presents some of the first solutions to this problem in the case of automatic speaker verification (ASV) systems.

First, the thesis reports an analysis of potential vulnerabilities and introduces an approach to evaluate ASV system performance in the face of spoofing. It presents the first comparison of established attacks (e.g. voice conversion and speech synthesis) and introduces a new threat in the form of non-speech signals (e.g. artificial signals). Also considered in the difference between spoofing attacks in terms of the effort required for their successful implementation. The thesis reports assessments with a number of ASV systems, from the standard GMM-UBM approach to the state-of-the-art i-vector scheme with PLDA post-processing. Experimental results show that all systems are vulnerable to spoofing. Voice conversion is the most effective attack and provokes increases in false acceptance rates to over 70%.

Second, the thesis presents three new spoofing countermeasures and their integration with state-of-the-art ASV systems. The first countermeasure is based on the detection of repetitive pattern which is effective in detecting artificial signals. The second is based on the analysis of feature dynamics which is effective in detecting converted voices. Like all competing approaches, both of these countermeasures make inappropriate use of prior knowledge. The third countermeasure therefore introduces for the first time the notion of generalized countermeasures, here implemented with one-class classifiers as a solution to outlier detection (unseen attacks). It exploits local binary pattern (LBP) analysis of speech spectrograms for feature extraction and one-class

support vector machine (SVM) classifiers. The generalised countermeasure, and therefore the most practically useful, achieves equal error rates (EER) of 5%, 0.1% and 0% in the detection of voice conversion, speech synthesis and artificial signal spoofing attacks respectively.

Résumé

En raison de la nécessité croissante pour les systèmes et les services sécurisés, la conception de systèmes fiables de reconnaissance personnelle devient de plus en plus importante. Dans ce contexte, les systèmes biométriques qui utilisent des traits physiologiques et/ou comportementales tels que les empreintes digitales, le visage, l'iris ou la voix pour la reconnaissance automatique des individus ont un certain nombre d'avantages par rapport aux méthodes d'authentification classiques tels que les puces, les cartes ou les passeports.

Cependant, en dépit des avantages, un nombre croissant de travaux indépendants montre que tous les systèmes biométriques sont vulnérables à la subversion, soit par le obscurcissement comme dans le cas de la surveillance ou, comme c'est l'intérêt de cette thèse, le leurrage (spoofing) dans le cas de l'authentification. étonnamment, il y a seulement très peu (mais à rythme croissant) de travaux visant à élaborer des contre-mesures qui peuvent offrir une certaine protection contre les attaques de type spoofing.

Cette thèse présente quelques-unes des premières solutions à ce problème dans le cas des systèmes de vérification automatique du locuteur (VAL).

Tout d'abord, la thèse fait état d'une analyse des vulnérabilités potentielles, et introduit une approche pour évaluer la performance du système VAL dans le scénario de l'usurpation d'identité. Elle présente la première comparaison des attaques établies (ex. la conversion de la voix et la synthèse de la parole) et introduit une nouvelle menace sous la forme de signaux non vocaux (signaux artificiels par exemple). Sont également considérés, la différence entre les attaques de type spoofing en terme de l'effort nécessaire pour leur mise en œuvre réussie. La thèse présente des évaluations avec un certain nombre de systèmes VAL, allant de l'approche standard GMM-UBM à l'état-de-l'art système du schéma i-vecteur avec post-traitement PLDA. Les résultats expérimentaux montrent que tous les systèmes sont vulnérables à les attaques de type spoofing. La conversion de la voix est l'attaque la plus efficace, et provoque une augmentation du taux de fausses acceptations à plus de 70%.

Deuxièmement, la thèse présente trois nouvelles contre-mesures et leur intégration dans les systèmes VAL de l'état de l'art. La première contre-mesure est basée sur la détection de motifs répétitifs qui est efficace pour détecter des signaux artificiels. La deuxième, quant à elle, est basée sur l'analyse de la dynamique de fond qui est efficace pour la détection de voix converties. Comme

toutes les approches concurrentes, ces deux contre-mesures font un usage inapproprié de la connaissance préalable. La troisième contre-mesure introduit donc pour la première fois la notion de contre-mesures généralisées, implémentées avec des classificateurs "1-classe", comme une solution pour détecter des attaques encore méconnues. Cette méthode exploite le motif binaire local (MBL) des spectrogrammes de la parole pour l'extraction des caractéristiques et une classe de la machine à vecteurs de support (MVS). La contre-mesure généralisée, et donc la plus utile dans la pratique, atteint des taux d'erreur (EER) de 5%, 0,1% et 0% dans la détection des attaques de type spoofing avec la conversion de la voix, la synthèse de la parole et les signaux artificiels respectivement.

“ I am still learning ”
(Michelangelo, age 87)

Acknowledgements

I would like to gratefully and sincerely thank my advisor Prof. Nicholas Evans for his countless hours dedicated to support, to encourage and to improve my research. Without his guidance and invaluable help this dissertation would not have been possible.

I wish to thank my committee members Professor John Mason, Professor Driss Matrouf, Professor Dirk Slock, Doctor Benoıt Fauve and Professor Artur Janicki for their precious time and for their constructive comments. Special thanks to Prof. Mason and Prof. Matrouf for the time dedicated to provide me a thorough review of my work.

I am also specially grateful to Dr. Fauve and Prof. Janicki, which not only have been traveled more than a thousand kilometers to attend my dissertation, but also have participated actively in my work and were of great help to improve it.

During my PhD I met many researchers which to a greater or lesser extent have influenced my work. The list is too long to be put here, but I would like to mention my colleague at EURECOM Ravichander Vipperla, which expertise in speech processing and computer skills were of great help at the beginning of my PhD; Prof. Sebastien Marcel and Prof. André Anjos for the discussions, advise and mails during our meetings during the project TABULA RASA and Prof. Tomi Kinnunen for his fruitful mails.

I would like to acknowledge and thank the secretaries, administration and in general the EURECOM institute for allowing me to conduct my research in an unbeatable environment and atmosphere, as well as the TABULA RASA members to share with me the adventure that this project was.

I would like to thank all my friends which made my life more enjoyable since my arrival at Nice (many to mention, but they know who they are). The moments I spent at EURECOM and in Cote d'Azur with this special people will always be in my memory.

Finally, I would like to extend my warmest thank to my family, who despite of the distance stood by me unconditionally through all thick and thin.

Contents

List of Figures

List of Tables

List of Publications

Papers published at conferences

(**C1**) F. Alegre, R. Vipperla, N. Evans and B. Fauve, "On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals", EUSIPCO 2012. Bucharest, Romania.

(**C2**) F. Alegre, R. Vipperla and N. Evans, "Spoofing countermeasures for the protection of automatic speaker recognition from attacks with artificial signals", INTERSPEECH 2012. Portland, Oregon, USA.

(**C3**) F. Alegre, A. Fillatre, N. Evans, "Spoofing countermeasures to protect automatic speaker verification from voice conversion", ICASSP 2013. Vancouver. Canada.

(**C4**) F. Alegre, R. Vipperla, A. Amehraye and N. Evans, "A new speaker verification spoofing countermeasure based on local binary patterns", INTERSPEECH 2013, Lyon, France.

(**C5**) F. Alegre, R. Vipperla, A. Amehraye and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns", BTAS 2013, Proceedings of the International Conference on Biometrics: Theory, Applications and Systems, Washington DC, USA.

(**C6**) F. Alegre, G. Soldi, N. Evans, B. Fauve and J. Liu, "Evasion and obfuscation in speaker recognition surveillance and forensics", IWBF 2014, 2nd International Workshop on Biometrics and Forensics, Valletta, Malta.

(**C7**) F. Alegre, G. Soldi and N. Evans, "Evasion and obfuscation in automatic speaker verification", ICASSP 2014. Florence, Italy.

(**C8**) F. Alegre, A. Janichi and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification", 2014 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany.

(**C9**) G. Soldi, S. Bozonnet, F. Alegre, C. Beaugeant and N. Evans, "Short-duration speaker modelling with phone adaptive training". The Speaker and Language Recognition Workshop ODYSSEY 2014. Joensuu, Finland.

Book chapters

(**B1**) N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre and P. De Leon "Speaker recognition anti-spoofing". Book Chapter in "Handbook of Biometric Anti-spoofing", Springer, S. Marcel, S. Li and M. Nixon, Eds., 2014

(**B2**) N. Evans, F. Alegre, Z. Wu and T. Kinnunen "Anti-spoofing: Voice Conversion" Book chapter in "Encyclopedia of Biometrics", Li, Stan Z. (Ed.), 2014

(**B3**) F. Alegre, N. Evans, T. Kinnunen, Z. Wu and J. Yamagishi "Anti-spoofing: voice databases" Book chapter in "Encyclopedia of Biometrics", Li, Stan Z. (Ed.), 2014

Journal

(**J1**) Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre and H. Li. "An overview of spoofing and countermeasures for automatic speaker verification". ELSEVIER Speech Communication Journal. 2014

TABULA RASA Deliverables[1]

(**D1**) D2.2: Specifications of biometric databases and systems. Submission date: 29/04/2011

(**D2**) D2.3: Specifications of spoofing attacks. Submission date: 31/07/2011

(**D3**) D2.5: Non-ICAO biometric databases of spoofing attacks. Submission date: 29/02/2012

(**D4**) D3.2: Evaluation of baseline non-ICAO biometric systems. Submission date: 30/09/2011

(**D5**) D3.4: Evaluation of baseline non-ICAO systems under spoofing attacks. Submission date: 30/07/2012

(**D6**) D4.2: Evaluation of initial non-ICAO countermeasures for spoofing attacks. Submission date: 31/12/2012

(**D7**) D4.4: Evaluation of advanced non-ICAO countermeasures for spoofing attacks. Submission date: 15/07/2013

[1]Deliverables (**D1**), (**D3**) and (**D4**) are publicly available at `https://www.tabularasa-euproject.org/project/deliverables`

Acronyms

Here are the main acronyms used in this document, sorted by alphabetical order

ACE	Average Classification Error
AS	artificial signals
asr	automatic speaker recognition space (GD-GMM)
ASV	automatic speaker verification
BMEC	Multi-modal Evaluation Campaign
CMS	cepstral mean subtraction
CSMAPLR	constrained structural maximum a posteriori linear regression
DET	standard detection error trade-off curves
DTW	dynamic time warping
EER	equal error rates
EM	expectation maximization
EPS	Expected Performance and Spoofability framework
FA	factor analysis
FAR	false acceptance rate
FFR	False Fake Rate
fil	filtering space (GD-GMM)
FLR	False Living Rate
FRR	false rejection rate
GD-GMM	Gaussian dependent filtering (voice conversion)
GLDS	generalized linear discriminant sequence kernel
GMM	Gaussian mixture model
GSL	supervector linear kernel
HIC	histogram intersection classifier
HMM	hidden Markov model
H-norm	zero handset normalisation
HTK	Hidden Markov Model Toolkit
HTS	Hidden Markov Model Synthesis Toolkit
ICAO	International Civil Aviation Organisation
IV	i-vector
JD-GMM	joint density gaussian mixture model (voice conversion)
JFA	joint factor analysis
LBP	local binary pattern
LDC	Linguistic Data Consortium
LP	linear prediction coefficients
LPCC	linear prediction cepstral coefficients

MAP	maximum a posteriori
MCS	Multiple Classifier Systems
MDC	mean distance classifier
MFCC	mel-frequency cepstral coefficient
ML	maximum likelihood
MLSA	Mel-logarithmic spectrum approximation filters
MSD-HSMM	multispace distribution hidden semi-Markov models
NAP	nuisance attribute projection
NIST	National Institute of Standards and Technology
OCC	one-class classification
OIC	with overlap index classifier
OPC	One per Class
PLDA	probabilistic linear discriminant analysis
PLP	perceptual linear prediction
PSOLA	pitch synchronous overlap add technique
PWD	pairwise distance analysis
RAPT	robust algorithm for pitch tracking
RPD	repetitive pattern detection
SAD	speech activity detection
SFAR	spoofing FAR
S-norm	symmetric score normalisation
SRE	Speaker Recognition Evaluation
SS	speech synthesis
SVM	support vector machine
T-norm	Test-normalization
TTS	text-to-speech
UBM	universal background model
VAD	voice activity detection
VC	voice conversion
VQ	vector quantization
WN	white noise
Z-norm	Zero-normalization

Chapter 1

Introduction

Biometrics refers to the technologies that measure and analyse a person's physiological and/or behavioural characteristics such as fingerprints, iris, voice, signature, face, DNA, gait, hand geometry and others for recognition (verification or identification) purposes. Biometrics offers an alternative over traditional methods for person recognition, relying on *what you are* or *what you do* as opposed to *what you know*, such as a PIN number or password, or *what you have* such as an ID card, a token or a passport.

The application areas in the field of biometric technology are vast and the number is growing. They include access control, border control, civil registry, entertainment, finance, forensic, health care, law enforcement, social media, social networking, surveillance, robotics, human-computer interaction, games, transportation, etc. The biometrics technology market is currently dominated by security-related applications and it is growing rapidly due to increasing security threats[1] in recent times.

The increase in unauthorized immigration, visa fraud, credit card fraud, border intrusion, and so on leads to a growing need for high security. Biometric technologies have been shown to be promising candidates for either replacing or augmenting conventional security technologies. For these biometric applications more than for others, *reliable recognition* is of great importance.

Reliable recognition relates to the insensitivity of biometric recognition systems against attempts to provoke a recognition error by interacting with them in a fraudulent manner. Although the state-of-the-art in biometric recognition has advanced rapidly in recent years, most of the efforts undertaken to develop this technology have been mainly directed to the improvement of recognition accuracy i.e. to lower recognition error rates, while reliability enhancement has been only partially addressed.

A growing bulk of independent efforts have shown that, independently of the modality, biometrics systems are affected by such threats. Surprisingly,

[1] `http://www.transparencymarketresearch.com/biometrics-technology-market.html`

there has been relatively little work in the development of countermeasures to protect such systems from the acknowledged threat of *subversion* [159].

This thesis presents some of the first solutions to this problem. In particular, it addresses some of the issues derived from a form of subversion denoted *spoofing* in order to trust automatic speaker verification (ASV) systems.

This thesis brings some insight into the problem of reliable recognition for ASV systems by the analysis and evaluation their vulnerabilities against spoofing, investigation of current (and possible new) threats and the development of new countermeasures that mitigate the effect of such threats.

1.1 Subversion

It is now well known that most biometric systems are vulnerable to some form of subversion. This section is reproduced from the author's own work previously published in [10].

Subversion aims to provoke a recognition error, either a false acceptance in the case of authentication applications, or a missed detection in the case of surveillance.

Surveillance applications typically involve the detection of one or more individuals, for example the detection of known criminals by mean of a closed-circuit television camera or in an intercepted telephone conversation. In such cases, persons of interest might disguise their biometric trait or manipulate their behaviour in order to *evade detection* [100, 155]. The intent here is to provoke a missed detection, otherwise referred to as a Type I error.

Authentication applications involve identification or verification scenarios in which an enrolled client typically seeks the confirmation of their identity in order to gain access to protected resources. The likely attack in this scenario involves *spoofing*, which entails the impersonation or manipulation of a biometric trait in order that it resembles that of an enrolled, target identity. The attack is thus intended to provoke a false acceptance, otherwise referred to as a Type II error.

There is arguably a third form of subversion, related more closely to traditional forensics, for example the analysis of DNA, fingerprints, hair samples, voice recordings etc. Here, biometric evidence can be manipulated, not only to evade reliable detection, but also so that they indicate the identity of another, specific person, i.e. to implicate another individual through the fabrication of

false evidence.

Reliable recognition performance is essential whatever the application. Spoofing can result in the granting of access to critical resources to persons of ill intent, whereas evasion and obfuscation can encumber or jeopardize criminal convictions. It is thus essential to accelerate the design of new approaches to detect manipulated traits and to ensure the reliability of automatic speaker recognition [10].

The work reported in this thesis relates to spoofing. Although it is not the main focus of this thesis, a study of evasion and obfuscation in the context of ASV is also presented in Appendix A.

1.2 Authentication & spoofing

Biometric security authentication systems present several advantages over classical authentication methods. In contrast to physical tokens or passwords, biometric information is generally not transferable in that it cannot be lost, forgotten, or guessed easily. Also there is nothing to remember or carry.

The system parameters of most biometric modalities can be tuned so they improve the authentication security[2] with respect to well accepted authentication methods i.e. PIN-based systems. Furthermore, the cost of integrating biometric components into an authentication system is continually decreasing, whereas the cost of relying on conventional authentication systems is increasing.

Biometric systems also presents a number of drawbacks, related mainly to the privacy issues due to the collection of biometric data and issues related biometric characteristics such as universality (an individual could not have hands) and permanence (most biometrics change over time), among others.

Other disadvantages are strictly related to security issues, such as the lack of secrecy (everyone knows our face or voice) and the fact that a biometric trait cannot be reset and/or replaced if compromised. Moreover, biometric security authentication systems are vulnerable to spoofing, which is the issue addressed in this thesis.

[2]Here the error of interest is false acceptance or Type II error

1.2.1 Motivation for spoofing countermeasures

Otherwise referred to as the direct, sensor-level or imposture attacks of biometric systems [69], *spoofing* refers to the presentation of a falsified or manipulated trait to the sensor of a biometric system in order to provoke a high score and illegitimate acceptance. Unless the biometric system is equipped with appropriate spoofing countermeasures, this threat is common to all biometric modalities. For example, face recognition systems can be spoofed with a photograph [63], whereas fingerprint or voice recognition systems can be spoofed with a fake, gummy finger [79] or with an audio recording [192], respectively.

Security systems must be constantly updated. A system that is assumed to be secure at the present day can become obsolete if it is not periodically improved. This is particularly true for biometrics, for which guaranteed reliability is a crucial requirement for the continued adoption of biometric systems in the security market.

While there is sufficient evidence that biometric systems are vulnerable to spoofing, it was not until recently that the research community started to address the problem actively.

1.2.2 Research initiatives

One of the earliest initiatives is the European TABULA RASA project[3], a pioneering study in the scientific community to address this issue. The goal was to research, develop and evaluate solutions to circumvent spoofing attacks, in order to increase trust in state-of-the-art biometric systems. The project considered biometrics adopted by standards (e.g. ICAO[4]) and also novel biometrics potentially more robust to spoofing. The consortium included six academic and six industrial partners. EURECOM's work involved 3D face and voice biometrics, the latter being the modality addressed in this thesis.

1.3 Voice biometrics & spoofing

From the number of different biometrics, the recognition of a person's identity using their voice has significant, wide-spread appeal; speech signals are

[3]The EU FP7 TABULA RASA project (`www.tabularasa-euproject.org`)

[4]The International Civil Aviation Organisation (ICAO) adopted face, fingerprint and iris biometric technologies.

readily captured in almost any environment using standard microphones and recording equipment, including remotely, i.e. over the telephone, where speech is usually the only biometric mode that is available. Since their natural appeal lies in automated, unattended scenarios, speaker recognition systems are particularly vulnerable to spoofing attacks.

Automatic speaker verification is a mature research field. However, in comparison to some other biometric modalities, spoofing and countermeasure research in ASV is far less advanced.

1.3.1 Spoofing in the context of ASV issues

This section explains subversion and in particular spoofing in terms of the variability of the input speech signal and into the context of current problems faced by speaker recognition.

The variability of the input signal (biometric trait) represents arguably the main adverse factor to accuracy in biometric systems. For voice modality, this issue is known as session variability [104] and refers to any variation between two recordings of the same speaker. Session variability is often described as mismatched training and test conditions, and it remains among the most challenging problems in speaker recognition.

Figure 1.1 illustrates different sources of variability for a speech signal. They can be grouped in variability due to changes in the acoustic environment and technical factors (transducer, channel) and those due to changes in a person's voice between two sessions. The latter can be divided in those variations which are unavoidable, intrinsic to the nature of the speech and *non-intentional* (state of health, mood, ageing) and those variations that include the *intentional* element, denoted in the biometric literature as subversion.

As mentioned before, while the scientific community has concentrated on mitigating the effects of the transmission channel and within-speaker (non-intentional) variability, a relatively more significant threat such as subversion is only just beginning to attract attention. Examples of ASV spoofing include impersonation, replay attacks, voice conversion and speech synthesis.

All of these approaches can be used to bias the distribution of impostor scores toward the true client or target distribution and thus to provoke significant increases in the false acceptance rate of ASV systems.

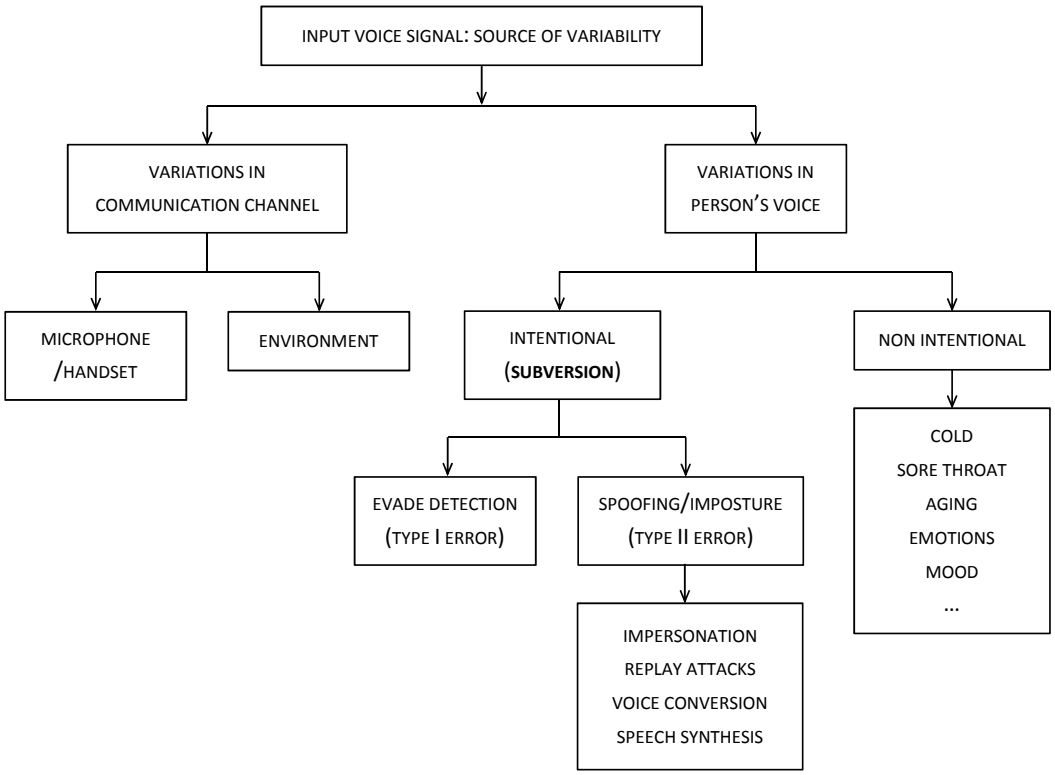


Figure 1.1: Categorization of voice variability, including both forms of subversion defined in Section 1.1.

1.3.2 Use cases

The market of automatic speaker recognition relates mainly to authentication, surveillance and forensic applications. Further details can be found in [24].

The voice modality is a non-ICAO biometric, which can be used for physical, logical and mobile access. It can be utilized standalone or as a part of multimodal configuration. It has particular appeal in mobile, remote access where recognition using voice and face (when the device is equipped with a camera) are natural choices of biometric.

Most of the use-cases can be grouped in either mobile/telephony or physical access scenario. Under the mobile/telephony scenario, we assume that the remote terminal is unknown (thus neither sensor nor channel are controlled) whereas for physical access scenario, both channel and sensor are controlled. Even though the attacks in the mobile/telephony scenario are not strictly aligned to the definition of "sensor level attack", in this thesis we consider them as direct attacks.

1.4 Research contributions

The following list shows research contributions in this PhD thesis. The list also refers to the published work, consisting in nine papers published at conferences (**C1**-**C9**), three book chapters (**B1**-**B3**) and one journal (**J1**) summarised in List of Publications.

- New literature review on spoofing and countermeasures
  - The author of this thesis has participated in the literature reviews included in (**B1**), (**B2**), (**J1**), (**D1**) and (**D2**). He has adapted them to be included in the Part I of this document.
- New spoofing attacks
  - A novel approach to artificial signals is first reported in (**C1**) and also in Section 4.2.1. To the best of our knowledge, this work is the first to consider the potential vulnerability of ASV to non-speech signals. Furthermore, experimental results reported in Section 5.3.2.2 show that attacks with artificial signals are a threat to ASV systems.
  - Experimental results reported in Section 5.3.2.2 suggest that attacks with low effort white noise signals are arguably a more serious threat compared to naive impostors.
- New insight on ASV system vulnerabilities, spoofing attacks and evaluations
  - An analysis of ASV vulnerabilities that complements the work in (**B1**) and (**J1**) is reported in Section 4.1
  - This work reassesses spoofing in terms of effort. It categorizes the threats in low, medium and high effort attacks together with the known zero-effort attacks (naive impostors).
  - Our experiments with white-noise together with some observations in the literature (i.e. the existence of the so-called *wolves* [35, 61] -impostor speakers that have natural potential to be confused with other speakers-) leads to the notion of generalized spoofing attacks (although the existence generalized spoofing attacks is not assessed in this thesis).
  - Chapter 5 reports the first study on vulnerability against spoofing with special focus in the effect of score normalization. The rela-

tion between score normalization and spoofing is also discussed in Section 5.3.3

  - The author addresses the problem of vulnerability evaluation for voice modality by a discussion presented in Section 4.3 and in (**B3**)

- New countermeasures

  - Three novel countermeasures are presented in this chapter. The first one relates with the detection of artificial signals and is presented in (**C2**), the second one is a specific approach to detect converted voices and is presented in (**C3**), the third one is a generalized approach with two variations, presented in (**C4**) and (**C5**), respectively.

  - This thesis report the first generalized approach to spoofing detection among all biometric modalities. This approach, based on one-class classification, is first presented in (**C5**) and also in (**J1**) and Section 7.4.

- New insight on the problem of reliable biometric verification

  - A novel, theoretical framework to address the problem reliable recognition is presented in Section 6.1. To the best of our knowledge, this work is the first to formulate spoofing and countermeasures as a multi-class problem with outliers detection.

  - The problem of countermeasure evaluation and integration is discussed in Section 6.2. In particular, the problem of countermeasure integration is addressed under the context of Multiple Classifier Systems (MCS).

- First countermeasures assessment in a multi-system, multi-attack framework

  - Previous studies on vulnerability evaluations against spoofing attacks are mostly under specific conditions i.e. one scenario, one spoofing attack and one ASV system. Presents the first comparative study with a wide family of ASV systems including the state-of-the-art i-vector scheme with PLDA post-processing.

  - Our work in (**C4**) is the first that evaluate countermeasures in a common multi-attack, multi-system framework for attacks of different nature including voice conversion, speech synthesis and artificial signals, respectively. The experimental part of this thesis

evaluate countermeasures for six different ASV systems including the state-of-the-art i-vector scheme with PLDA post-processing and the three mentioned attacks.

- First study of evasion and obfuscation with large-scale standard datasets
  - This thesis is the first in reassessing the problem of obfuscation by the classification and study of independent attacks perpetrated at the biometry detection level and at the recognition level which are redefined as evasion and obfuscation, respectively. This work is first reported in (**C7**).
  - The work reported in (**C6**) and (**C7**) are the first to report results on obfuscation of large-scale, standard (NIST) databases.
  - The work reported in (**C7**) shows that LBP-based countermeasure shows can detect evasion and obfuscation with reasonable accuracy.

Other contributions developed within the framework of this thesis but not included in this document:

- First replay attack assessment with large-scale standard datasets
  - Our work in (**C8**) reassess the threat of replay attacks. Results shows that, despite the lack of attention to replay attacks in the literature, low-effort replay attacks pose a significant risk, surpassing that of comparatively high-effort attacks such as voice conversion and speech synthesis.
- First comparative study of attacks for physical-access scenario
  - EURECOM database: a MOBIO-style database consisting approximately 18 hours of audio/video samples from 21 subjects (1260 samples from 14 males and 7 females) was collected to evaluate spoofing under physical access scenarios. Details and protocols are presented in (**D3**) .
  - A GMM-UBM based ASV system is evaluated for the physical access scenario (i.e. EURECOM database) and for four different attacks, including replay attack, voice conversion, speech synthesis and artificial signals. Results are presented in deliverables (**D4**) to (**D7**).
- New work on countermeasures

- The work reported in (**C2**) shows the first study that utilized ITU-T specifications as speech quality assessment countermeasure, in this case against attacks with artificial signals. Against intuition, results are not satisfactory, which opens the discussion of the need of dedicated effort for countermeasure development.
- A countermeasure based on one-dimensional local binary patterns (LPB) is presented in (**D6**).
- The first experimental work that shows results related to the fusion of two countermeasures is reported in (**D6**)

1.5 Outline of the thesis

The work reported in this thesis is divided in two parts. Part I summarizes the state-of-the-art in speaker recognition, spoofing and countermeasures and provides a theoretical support to the contents included in the second part of the thesis. Part II describes the new contributions.

The structure is as follows:

- Chapter 1 introduces the problem and describes the motivation, outline and contributions of this PhD thesis.
- Chapter 2 summarizes the work in speaker recognition, including the description of state-of-the-art ASV systems, evaluation databases, protocols, metrics and software packages and platforms.
- Chapter 3 summarizes related work in ASV vulnerability assessment and countermeasure development. It focuses mainly in the description of spoofing attacks by impersonation, replay, speech synthesis and voice conversion and corresponding countermeasures.
- Chapter 4 reports our assessment of ASV system vulnerabilities and spoofing attacks. In particular, this chapter present an analysis of potential vulnerabilities in ASV systems with respect to spoofing attacks, it reassess the problem of spoofing by introducing new threats in the form of non-speech audio signals and also accesses spoofing attacks in terms of effort. Finally, it discusses common issues related to the vulnerability evaluation of ASV systems.
- Chapter 5 reports a comparative study of the effectiveness (risk) of the spoofing attacks described in Chapter 3- 4 for a number of ASV systems.

The chapter includes the specifications of ASV systems and spoofing attacks, biometric and spoofing databases, protocols and metrics adopted for each evaluation.

- Chapter 6 discusses some fundamental issues before addressing countermeasure development. First, this chapter reports a thorough analysis of the problem of speaker recognition under the new paradigm of spoofing attacks and countermeasures. Second, it reports our analysis of countermeasure integration together with some weaknesses in current evaluation methodologies.
- Chapter 7 introduces three novel countermeasures proposed in the framework of this thesis that are used later in the experimental part. The first countermeasure is based on feature distribution analysis which is effective in detecting artificial signals. The second is based on a pairwise distance analysis to detect converted voices and the third uses local binary patterns for generalized attack detection.
- Chapter 8 extends the evaluation reported in Chapter 5 by adding the evaluation introduced in Chapter 7. The countermeasures are evaluated stand-alone and also integrated to the ASV systems. The chapter includes countermeasures description and setup and protocols and metric adopted for each evaluation.
- Chapter 9 concludes the thesis summarizing the main results obtained and outlining future research directions.
- Appendix A Reports our analysis and experimental work for evasion and obfuscation in speaker recognition systems.

Readers expert in the field of speaker recognition may avoid Chapter 2 and use Chapter 3 for consultation. The dependency among the chapters in Part II is linear i.e. starting from Chapter 4 until Chapter 9.

Part I

LITERATURE REVIEW

Chapter 2

Automatic speaker recognition

With the growth in telecommunications and vast related research effort, voice-based authentication has emerged over the last decade as a popular and viable biometric. Speaker recognition is generally the preferred or even only mode of remote verification over the telephone, for example. Speaker recognition also has obvious utility in multi-modal biometric systems where it is commonly used with face recognition.

Some of the first work in speaker recognition was reported in the 1970s [17] and developments since then can be traced in [62, 35, 24]. Speaker recognition is today an extremely active and mature field of biometrics research.

Speaker recognition systems are either text-dependent or text-independent with the latter having received the greatest attention in the open literature. An appropriate measure of today's state-of-the-art can be found in the proceedings of the internationally competitive Speaker Recognition Evaluations (SREs) [142] that are administered by NIST in the US. Through these evaluations, running since 1997, researchers have been able to reliably compare different approaches using common experimental protocols and large datasets; this alone has facilitated much of the progress made in the field over recent years.

Historically, the standard approach to text-independent speaker verification involves cepstral-based features and Gaussian mixture models (GMMs). Some of the main developments have involved the use of support vector machine classifiers, various feature, model and score normalization approaches and, more specifically, channel normalization/compensation and/or both intra-speaker and inter-speaker variability modelling which led to the state-of-the-art i-vector scheme. These technologies are responsible for some of the most significant advances over recent years and have evolved into a core focus of the research community.

This chapter describes the state-of-the-art in ASV, including the description of state-of-the-art ASV systems, evaluation databases, protocols, metrics and software packages and platforms, making emphasis on the technologies, tools

and evaluation methodologies used in this thesis.

The review presented in this chapter is not exhaustive. The contents are adapted to facilitate the comprehension of the different ideas introduced along this thesis. More general and detailed overviews of the fundamentals can be found in [35, 24, 105, 117, 186, 116].

2.1 Fundamentals

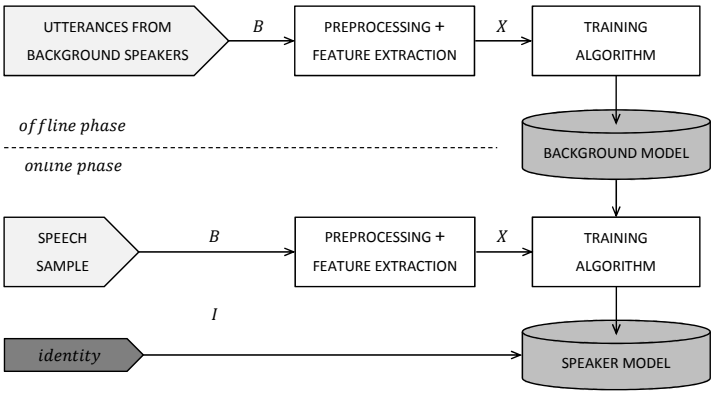
A schematic representation of a generic speaker recognition system architecture is presented in Figure 2.1. In the enrolment phase, illustrated in Figure 2.1(a), a speaker model is trained with the help of a set background speakers, which are used either as the negative examples in the training of a discriminative model [38], or in the training of a universal background model (UBM) which represents the alternative hypothesis in statistical modelling [162].

Depending on the context, a speaker recognition system can be used either in a verification mode or an identification mode. In verification mode (Figure 2.1(b)), a person's claimed identity is confirmed based upon validating a collected utterance against the model of that individual. On the other hand, in identification mode (Figure 2.1(c)), the system has to recognize a person based upon the comparison of a collected utterance against a collection of models of N individuals.

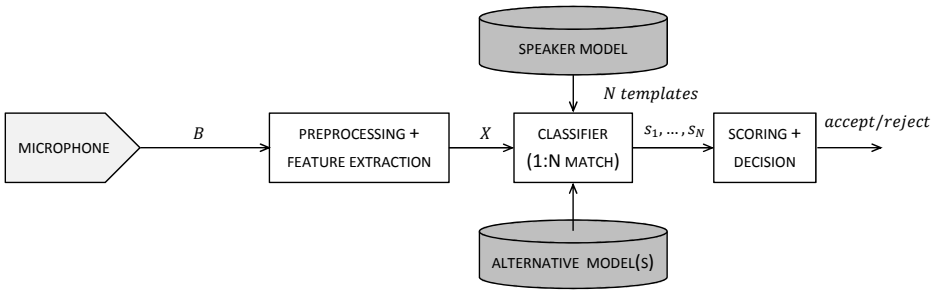
There are two general modes of speaker recognition. They are text-dependent and text-independent. In text-dependent systems [88], suited for cooperative users, the recognition phrases are fixed, prompted or known beforehand. In text-independent systems, there are no constraints on the words which the speakers are allowed to use. The majority of speaker recognition research relates to text-independent speaker verification and is the focus in this thesis.

2.2 ASV systems

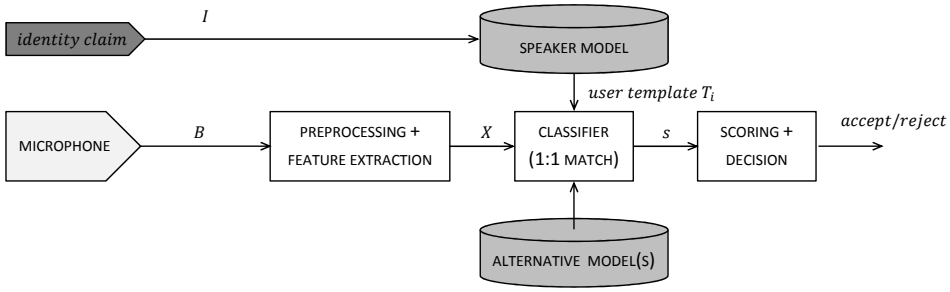
This section describes state-of-the-art approaches to text-independent automatic speaker verification (ASV) by means of the analysis of the modules presented in Figure 2.1 and by a brief review on fusion of ASV systems.



(a) Speaker enrolment



(b) Speaker identification



(c) Speaker verification

Figure 2.1: Block diagram of a typical speaker recognition system, consisting in enrolment followed by identification/verification modes. Figure reproduced from [77].

2.2.1 Preprocessing

The preprocessing module involves operations such as detection of the pattern of interest from the background, noise removal and pattern normalization,

among others.

Detection aims to identify those components or intervals of the input signal which are of interest to the recognition module, i.e. typically the components containing a face, a fingerprint, or intervals containing speech [21]. Detection is arguably the most important preprocessing step in a recognition system and a sub-component present in any real-world implementation, and thus the one addressed in this thesis.

In terms of ASV, biometry detection is commonly referred to as either speech activity detection (SAD) or voice activity detection (VAD). Three predominant forms are used in practice but, be they energy-based, model-based of phoneme-based detectors, the goal is common to all, namely to identify frames in the input signal which contain useful speech.

The most simple energy-based SADs are still well accepted in the literature and, mostly for reasons of computational simplicity, they might be preferred in practice. There is also some recent evidence [172] which suggests energy-based SAD can be more effective for ASV applications than model and phoneme-based SADs and also that standardised in G.729B [21] for both clean conditions and different types of noise.

2.2.2 Feature extraction

Most state-of-the-art speaker verification systems use features that are based on short-term spectral estimates i.e. short-term segments (frames) of 20 to 30 msec in duration. Typically, mel-frequency cepstral coefficient (MFCC), linear predictive cepstral coefficient (LPCC) or perceptual linear prediction (PLP) features are used as a descriptor of the short-term power spectrum. These are usually appended with their time derivative coefficients (deltas and double-deltas) and log-energy.

Also various feature-level normalization approaches have been investigated. Mostly aimed at attenuating channel effects, they include cepstral mean subtraction (CMS) [75], RASTRA filtering [91], feature warping [151] and feature mapping [161].

In addition to spectral features, prosodic and high-level features have been studied extensively [176, 58, 177], achieving comparable results to state-of-the-art spectral recognizers [109]. Prosodic features are extracted from longer segments such as syllables and word-like units to characterise speaking style and intonation, while high-level features aim to represent speaker behaviour

or lexical cues.

For more details regarding popular feature representations used in ASV, readers are referred to [105].

2.2.3 Modelling & classification

Most of the approaches to text-independent ASV have their roots in the standard GMM with a universal background model, the so-called GMM-UBM approach [160, 162, 24]. The most common implementation utilizes a UBM which is trained using expectation maximization (EM) and large amounts of data from a pool of background speakers. Due to the common lack of speaker-specific data, target speaker models are generally adapted from the UBM using maximum a posteriori (MAP) adaptation [83]. Scores correspond to normalised log-likelihood ratios computed from the target and background models. Additional normalization strategies operating at the score level are described in Section 2.2.4.

The state-of-the-art has advanced significantly since the early days of GMM-based approaches. Support vector machines (SVMs) [191] have become a popular approach to pattern classification and speaker verification is no exception. Early attempts to use SVMs for speaker verification appeared in the mid-to-late 90's e.g. [174, 37]. These early approaches used cepstral-based parameterisations and led to results that were inferior to a standard GMM.

More recent SVM-based approaches such as the generalized linear discriminant sequence kernel (GLDS) [38] and the GMM supervector linear kernel (GSL) [39] approaches are capable of outperforming the standard generative GMM-based approach [71]. The GSL approach is one example where the input to the SVM classifier comes from a conventional GMM and is here the concatenation of the GMM mean vectors [39] better known as the GMM supervector.

Despite harnessing the discriminative power of the SVM the above approaches do not explicitly model inter-session variability which the next generation of speaker verification system sought to achieve. There have been two main approaches, namely nuisance attribute projection (NAP) [178] and joint factor analysis (JFA) [104].

The NAP approach aims to attenuate session effects in a discriminative SVM framework. JFA has received a huge amount of attention and there are numerous implementations reported in the literature, e.g. [104, 194, 135]. In

contrast to feature mapping [161] the JFA approach assumes that the channel variability space is continuous instead of discrete and combines a model of both speaker and session variability.

Joint factor analysis approaches have proved to be among the best performing approaches to date, but by mean of a tedious process used to train the speaker and session subspace models. Subsequently, JFA evolved into a much-simplified model that is now the state-of-the-art. The so called total variability model or 'i-vector' representation [59] uses latent variable vectors of low-dimension (typically 200 to 600) to represent an arbitrary utterance.

After the i-vector extraction step, post-processing techniques are applied to attenuate session effects. In particular, probabilistic linear discriminant analysis (PLDA) [118] with length-normalised i-vectors [81] has proven particularly effective.

Further details on systems and modelling techniques used in this thesis are described in Section 2.3.3.2. For more details regarding modelling and classification used in ASV, readers are referred to [105].

2.2.4 Scoring & decision

Score normalization is part of the state-of-the-art of a wide family of GMM-based speaker recognition systems. Other than some i-vector schemes, which seems not to need them [102], score normalization techniques have been used in most of the research to improve recognition performance at the expense of computational cost.

Let $L_\lambda(X)$ denote the score for speech signal X and speaker model λ. The normalised score $\widetilde{L}_\lambda(X)$ is then given as follows:

$$\widetilde{L}_\lambda(X) = \frac{L_\lambda(X) - \mu_\lambda}{\sigma_\lambda} \tag{2.1}$$

where μ_λ and σ_λ are the estimated mean and standard deviation of the impostor score distribution, respectively.

Among the various normalization techniques, Zero-normalization (Z-norm) and Test-normalization (T-norm) are the most widely used methods to estimate the normalization parameters, μ_λ and σ_λ.

In Z-norm, during the training stage a set of impostor utterances is scored against each potential claimant model while in T-norm [18] during the test

stage the test utterance is scored against a pre-selected set of cohort models (preselection based on the claimant model). In both cases the resulting score distribution is then used to estimate the normalization parameters in Equation 2.1.

The advantage of T-norm over Z-norm is that any acoustic or session mismatch between test and impostor utterances is reduced. However, the disadvantage of T-norm is the additional test stage computation in scoring the cohort models [19].

Additional strategies include handset-dependent normalisation such as zero handset normalisation (H-norm) [64] and handset-dependent T-norm (HT-norm). For detailed information readers are referred to [24].

2.2.5 System fusion

Although it is not explicitly represented in Figure 2.1, any state-of-the-art review would be complete without referring to the many attempts to bring additional improvements in performance through the fusion of different systems and scoring approaches, some notable examples including [135, 57, 85]. An excellent comparison of these approaches using common parameterisations and datasets is presented in [71, 72].

Opposite to most of this work, this thesis necessarily focuses only in fusion techniques to enhance robustness instead of recognition performance. Fusion is a key component within the Multiple Classifiers Systems (MCS) theory. MCS is treated in detail in Section 6.2 as a part of the contribution of this thesis related to countermeasures integration to ASV systems.

2.3 Datasets, protocols, metrics & software

The following section summarizes the literature in speaker verification related to existing datasets, protocols, metrics and publicly available software packages. The contents introduced in this section are adapted from the author's own work previously published in TABULA RASA deliverables D2.2 and D3.2 [15].

2.3.1 Databases & protocols

As with any biometric, speech databases suited to the speaker verification task should in general have a large, representative number of speakers. Since speech characteristics from the same person can vary significantly from one recording to another there is a requirement for multi-session data which should reflect differences in acoustic characteristics related not only to the speaker, but also to differing recording conditions and microphones.

It has been suggested that collection over a period of three months [76] is a minimum in order to reliably capture variations in health and fatigue for example.

Since the information in a speech signal is contained in its variation across time, i.e. it is a dynamic signal, speaker verification performance also varies significantly depending on the quantity of data used both for training and testing.

2.3.1.1 Existing databases

Databases such as TIMIT [82], Aurora [150] and Switchboard [86] are all used widely for speech technology research. Though these databases have also been used for speaker recognition evaluations to some extent, they are designed primarily for speech recognition research. In addition these corpora are somewhat limited in the variety of microphone and recording conditions and also well defined development, evaluations, training and testing subsets targeted for speaker recognition experimentation.

Existing corpora, such as the CHAINS [50], YOHO [36] and CSLU [49] datasets are specific to speaker recognition, but have a limited number of speakers. The EVALITA [20] evaluations provided for a dedicated speaker recognition task in 2009 but did not feature in 2007 and is not included in the evaluation plan for 2011.

The Speaker Recognition Evaluation (SRE) [129] datasets collected by the National Institute of Standards and Technology (NIST) are presently the de facto standard evaluation platform for speaker recognition research and are the only realistic means of gauging the state-of-the-art. Since they provide for large, multi-session datasets with different evaluation conditions and since they facilitate comparisons to the existing state-of-the-art, **the NIST SRE datasets are used for speaker verification work in this thesis**.

Since not all NIST datasets are publicly available we have decided to restrict those used in this thesis to datasets that are or become publicly available during the period of the thesis (2011-2014) through the Linguistic Data Consortium[1] (LDC). We also note that the same datasets have been used previously in related work [29].

Finally we refer to two multi-modal datasets that include a speech component. The three datasets from the BioSecure Multi-modal Evaluation Campaign (BMEC) [148] contain seven different modes including speech. The MOBIO dataset [126] contains both face and voice modalities and are used in TABULA RASA for 2D-face and voice.

2.3.1.2 The NIST SRE datasets

The 2003, 2004, 2006 and 2008 NIST SRE datasets are currently publicly available through the LDC. They contain several hundreds of hours of speech data collected over the telephone including some calls made using mobile telephones. Further details of each dataset are available from the LDC website with additional information available from the NIST SRE website[2].

Each evaluation involves one compulsory, 'core' experiment and several other optional experiments. The differences between each experiment or condition entail mostly different quantities of training and/or test data and possibly varying channel conditions. Training and testing protocols are defined and allow for different systems and technologies to be readily and meaningfully compared according to standard experimental and evaluation protocols and metrics.

A typical speaker recognition system requires an independent development set in addition to independent auxiliary data which is needed for background model training and the learning of normalisation strategies. This data typically comes from other NIST datasets, such as the 2004 dataset and is the case for all the work in this thesis. Also, NIST SRE 2008 dataset is used to cope with the data needed for baseline systems based on i-vectors. All NIST SRE datasets have a very similar specification. Full details about NIST'05 and NIST'06, used for development and evaluation respectively, can be found in their respective evaluation plans [143, 144] and also in [156].

[1] `http://www.ldc.upenn.edu/`

[2] `http://www.itl.nist.gov/iad/mig/tests/sre`

2.3.2 Evaluation metrics

The evaluation of ASV systems requires large numbers of two distinct tests: target tests, where the speaker matches the claimed identity, and impostor tests, where the identities differ. Accordingly, the ASV system is required to either accept or reject the identity claim, thereby resulting in one of four possible outcomes, as illustrated in Table 2.1. There are two possible correct outcomes and two possible incorrect outcomes, namely false acceptance (or false alarm) and false rejection (or miss).

	Accept	Reject
Genuine	Correct acceptance	False rejection
Impostor	False acceptance	Correct rejection

Table 2.1: Four categories of trial decisions in automatic speaker verification.

Statistics acquired from many independent tests (trials) are used to estimate the false acceptance rate (FAR) and the false rejection rate (FRR). The FAR and FRR are complementary in the sense that, for a variable threshold and otherwise fixed system, one can only be reduced at the expense of increasing the other.

In practice, all system parameters are optimised to minimise the balance between FAR and FRR, which is commonly measured in terms of the equal error rate (EER)[3], although this is certainly not the only optimisation criterion. Another common metric present for core NIST evaluations is defined as follows:

$$C_{Norm} = \frac{C_{Miss} \times P_{Miss/Target} \times P_{Target} + C_{FA} \times P_{FA/NonTarget} \times P_{NonTarget}}{C_{Default}} \tag{2.2}$$

where the cost of a miss and of a false alarm (FA) are 10 and 1 respectively, where the probability of a target and non-target are 0.01 and 0.99 respectively and where the normalisation factor $C_{default} = 0.1$ is defined in order that a system which always returns a negative decision obtains a score of $C_{Norm} = 1$.

While the C_{Norm} metric is the default, dynamic performance, including a comparison of minimum and actual costs (i.e. with regard to optimised and actual thresholds), are compared according to standard detection error trade-off (DET) curves [128].

[3]EER corresponds to the operating point at which FAR=FRR.

2.3.3 Platforms & software packages

Given the complexity of standard NIST speaker recognition datasets it is desirable that the system adopted is well-adapted and suited to running such evaluations. Thus, computational efficiency is also a requirement. Speaker recognition experiments can also be performed in a multi-modal setting and thus it is also sensible that the system may be used in conjunction, or fused with a face recognition system. In the following we review some existing tools that are appropriate in this case and then describe in more detail the system adopted for this thesis.

2.3.3.1 Existing tools

Speaker recognition systems have advanced rapidly over the last few decades and there exist some useful software packages and libraries that can be used to build state-of-the-art speaker recognition systems with relative ease.

SPro[4], the open-source speech signal processing toolkit, provides for highly configurable feature extraction. The Hidden Markov Model Toolkit (HTK)[5] and the Hidden Markov Model Synthesis Toolkit (HTS)[6] provide a set of tools for building statistical speaker models and can also be used for feature extraction. Matlab[7] from Mathworks Inc. has various toolkits for statistical pattern recognition and is an excellent tool to prototype quickly a speaker recognition system and to develop advanced algorithms. In this regard, the MSR Identity Toolbox[8] from Microsoft Research has gained some popularity due of its simplicity to quickly build state-of-the-art (i.e. i-vectors + PLDA) baseline systems. Octave[9], its open source equivalent, also provides powerful features.

The ALIZE/Mistral platform[10] [32, 138] is a library for biometric authentication and provides a comprehensive set of functions related to the task of statistical speaker recognition. LIA-RAL[11] is a set of tools for speaker recognition and is built using the ALIZE/Mistral library. libsvm[12] is a library which

[4] http://www.gforge.inria.fr/projects/spro
[5] http://htk.eng.cam.ac.uk/
[6] http://hts.sp.nitech.ac.jp/
[7] http://www.mathworks.com/products/matlab/
[8] http://research.microsoft.com/apps/pubs/default.aspx?id=205119
[9] http://www.gnu.org/software/octave/
[10] http://www.lia.univ-avignon.fr/heberges/ALIZE/
[11] http://www.lia.univ-avignon.fr/heberges/ALIZE/LIA_RAL
[12] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

provides for support vector classification and has been integrated into LIA-RAL. The Torch toolkit[13] also has robust implementation of Support vector machine based classifiers. Finally FoCal[14], a set of Matlab functions for the fusion and calibration of multiple classifiers, has proven very popular in the speaker recognition community.

SPro, ALIZE, LIA-RAL and FoCal are arguably the most popular tools for speaker recognition. They are all open-source, are used in combination by many independent teams and have achieved state-of-the-art performance in the NIST speaker recognition evaluations. Furthermore the ALIZE/MISTRAL toolkits have been used for voice transformation in order to demonstrate the threat from spoofing. This combination was used for all work in this thesis.

2.3.3.2 The ALIZE speaker recognition system

The 'ALIZE' speaker recognition system is something of a misnomer since ALIZE is really a library, not a toolkit. Even so, the open-source LIA-RAL toolkit, which does provide a set of executable for speaker recognition, has inherited the name of the library on which it is based. In this thesis, work in speaker recognition will be based upon the implementation described in [72].

While ALIZE has native support for most standard feature file formats, SPro is the most popular. SPro provides for both Mel [139] and linear scaled frequency cepstral coefficients in addition to linear prediction coefficients, static and dynamic features.

Features generally encompass some channel characteristics which manifests as convolutional noise. Under conditions of mismatched training and testing these effects can lead to significant degradations in performance and some means of channel compensation generally prove beneficial. In the standard baseline setup this includes cepstral mean and variance normalisation.

ALIZE also provides a comprehensive suite of different feature normalisation strategies including feature warping [151], feature mapping [161] and factor analysis eigen-channel compensation [103].

The standard approach to statistical speaker modelling is based on Gaussian mixture models (GMMs) [162] and is the approach adopted in ALIZE. First, a world model [41] or universal background model (UBM) is trained using

[13] `http://www.idiap.ch/scientific-research/resources/torch`
[14] `http://sites.google.com/site/nikobrummer/focal`

expectation maximisation (EM) [60] and large amounts of data from a pool of background speakers. Due to the common lack of speaker-specific data, target speaker models are generally adapted from the UBM during enrolment through maximum a posteriori (MAP) adaptation [83]. Although all the parameters of the UBM can be adapted, the adaptation of the means only has been found to work well in practice [162] and is the approach in the largely standard baseline system.

The ALIZE framework also provides for more recent approaches which harness the power of SVMs, joint factor analysis (JFA) and i-vectors (the latter included in ALIZE 2.0 and subsequent versions). Support vector machines (SVMs) [190] have become a popular approach to pattern classification and speaker verification is no exception. The more recent SVM-based approaches such as the generalised linear discriminant sequence kernel (GLDS) [38] and the GMM super-vector linear kernel (GSL) [39] are capable of outperforming the standard GMM-based approach and are supported in ALIZE. The GSL approach is one example where the input to the SVM classifier comes from a conventional GMM and is formed from the concatenation of the GMM mean vectors into the so-called GMM super-vector [39].

Other approaches supported in ALIZE include nuisance attribute projection (NAP) [40], joint factor analysis (JFA) [101] and an i-vector (IV) [59] extractor plus back-end techniques such as cosine similarity, mahalanobis distance, two-covariance modelling [34] and probabilistic linear discriminant analysis (PLDA) [118], the latter being the approach chosen for i-vector related experiments presented in this thesis.

Chapter 3

Spoofing & countermeasures

In the past few years, a considerable effort has been carried out in analysing, classifying and solving the possible security breaches that biometric verification systems may present. For voice, while earlier work considered the threat from classical spoofing attacks such as impersonation [28, 68] or replay [119, 192], that from more advanced attacks has attracted attention only recently. Attacks from voice conversion [152, 153, 133, 106] and speech synthesis [130, 54] have all been shown to provoke significant increases in the false acceptance rate of state-of-the-art ASV systems.

Reassuringly, as has been the case for other biometric modalities, e.g. face recognition [122, 43, 15], the speaker recognition community has started to address the problem through efforts to develop specific spoofing countermeasures [145, 53, 56, 198, 200, 12]. However, in comparison to some other biometric modalities, spoofing and countermeasure research in ASV is far less advanced.

This chapter reviews related work in ASV vulnerability assessment and countermeasure development. Section 3.1 aims to define some of the terminology used in this thesis and to put in context spoofing and countermeasures for voice helped by the related literature on other biometrics. Section 3.2 reviews past work to evaluate vulnerabilities and to develop spoofing countermeasures. We consider impersonation, replay, speech synthesis and voice conversion. Finally, Section 3.3 summarizes the current approaches to evaluate spoofing and countermeasures.

3.1 Definitions & assumptions

Spoofing attacks are performed on a biometric system at the sensor or acquisition level to bias score distributions toward those of genuine clients, thus provoking increases in the false acceptance rate (FAR). Two key elements are characteristic of a spoofing attempt; first, the presence of malicious client or subject, from now on called the *spoofer*, and second a non-zero probability of

the perpetrated attack to succeed in fooling the biometric system.

The classification of vulnerabilities and countermeasures followed in this thesis is based on the general scheme presented in [77], but adapted for the voice modality.

3.1.1 Attacks

In general, the attacks that can compromise the security provided by a biometric system are categorized into two basic types, denoted brute-force and adversary attacks [94]. To be in line with the concept of *effort* later introduced in Section 4.2, in this thesis brute-force and adversary attacks are denoted **zero-effort attacks** and **nonzero-effort attacks**, respectively (Figure 3.1).

Vulnerabilities to brute-force or zero-effort attacks, also denoted as intrinsic failure [93], are present in all biometric systems and are impossible to prevent. They are derived from the fact that there is always a probability that two speech samples coming from two different speakers are sufficiently alike to produce a positive match. With this type of attack the impostor uses the systems in a normal and straightforward manner with the hope of overcoming the system by chance.

Adversary attacks or nonzero-effort-attacks refer to the possibility that a malicious subject, enrolled or not to the application, tries to bypass the system by interacting with it in a fraudulent manner e.g. hacking an internal module, using a fake biometric trait, deliberately manipulating his biometric trait to avoid detection, etc.

Zero-effort vulnerabilities are inherent to the statistical nature of biometric systems and are already addressed in the conventional biometric recognition problem. Hence, the biometric community has focused in the development of specific countermeasures against adversary attacks. These attacks have been identified by [158] and categorized depending on the point at which the attack is directed. They are illustrated in Figure 3.2 for the voice modality.

Two types of attacks are broadly considered: direct attacks and indirect attacks. **Direct attacks** are traditionally related to the attacks performed at the sensor level (point 1 in Figure 3.2). Here, an adversary, typically referred to as an impostor, might seek to deceive the system by impersonating another, enrolled user at the sensor or microphone level in order to manipulate the ASV result. In the specific case of ASV, attacks at both the sensor and transmission levels are generally considered to pose the greatest threat [70]. Thus, spoof-

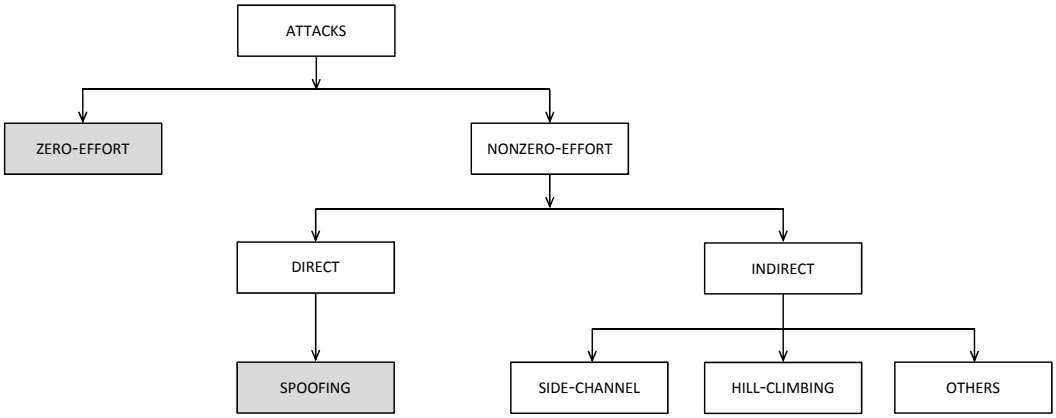


Figure 3.1: Classification of attacks (adapted from [77] for voice modality). Zero-effort attacks and spoofing attacks, which are analysed in the experimental part of the thesis, are illustrated by shaded boxes.

ing attacks considered in this thesis also include transmission-level attacks i.e. attacks where speech signals are intercepted and replaced at the transmission level by another specially crafted voice signal (point 2 in Figure 3.2).

Indirect attacks are performed inside the system and are due to intruders, such as cyber-criminal hackers, by bypassing the feature extractor or the matcher (points 3 and 5 in Figure 3.2), by manipulating the templates (or models) in the database (point 6 in Figure 3.2), or by exploiting the possible weak points in the communication channels (points 4, 7 and 8 in Figure 3.2). Most of the work regarding indirect attacks use some type of variant of the hill-climbing technique introduced by [180], although recently the so-called side-channel attacks, arguably a bigger threat for biometric systems than hill-climbing approaches, are receiving increased attention [78].

This thesis focuses on spoofing i.e. on attacks at points 1 and 2 in Figure 3.2. Zero-effort attacks are not considered spoofing attempts (although they meet the two requirements defined in Section 3.1 to be considered spoofing attacks). Nevertheless, zero-effort attacks are part of the ASV performance evaluations and they are thus included in the experimental work in this thesis.

Attacks perpetrated during the enrolment process are not considered in this thesis. They can be either considered as indirect attacks (attacks at point 6 in Figure 3.2) or the enrolment process can be assumed secure. In any case, they will not be treated any further here.

A list of threats not addressed in this thesis that may affect any security application, not only based on biometric recognition, is presented in [123].

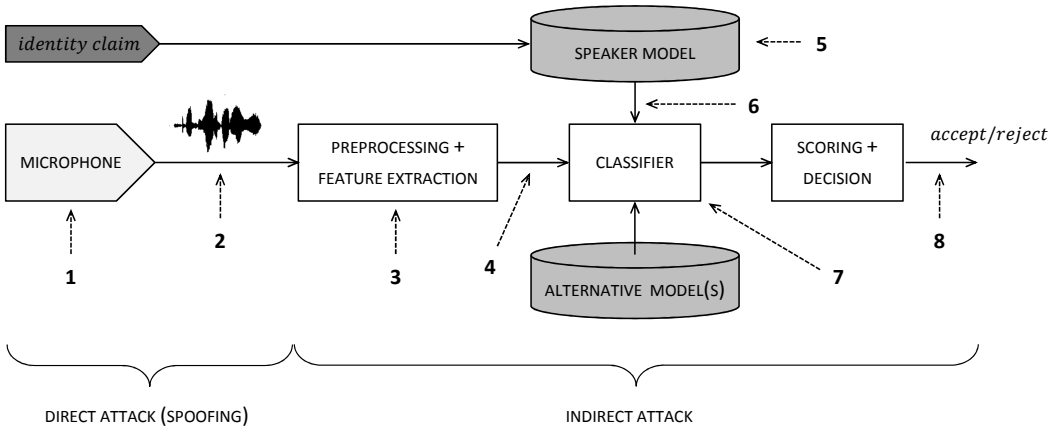


Figure 3.2: An illustration of a typical ASV system's verification mode (same as Figure 2.1(c)) with eight possible attack points. Attacks at points 1-2 are considered as direct attacks whereas those at points 3-8 are indirect attacks. The figure is adapted from [158, 199] for voice modality)

It includes circumvention (an unauthorized user gains access to the system), collusion (a user with special privileges, e.g. an administrator, allows the attacker to bypass the recognition component) and coercion (legitimate users are forced to help the attacker enter the system), among others.

3.1.2 Attack protection

One possible way to categorize the known methods to minimize the risk arising from attacks at the sensor and transmission level is presented in Figure 3.3. Following the scheme in [77], the biometric-based attack protection method can be divided into preventive and palliative approaches.

Preventive methods aim to avoid a certain attack to be perpetrated. They include mostly security measures that offer specific protection from templates [2, 42, 93, 189]. In particular, previous work [95, 206] shows that *watermarking*, where extra information is embedded into the host data, can be useful to protect systems where the sensor and the transmission channel are not secured.

Palliative methods aim to minimize the risk of an attack breaching the system after the attack has been produced. Among other palliative countermeasures to direct attacks, **liveness detection** approaches which have received the greatest attention from researches and industry.

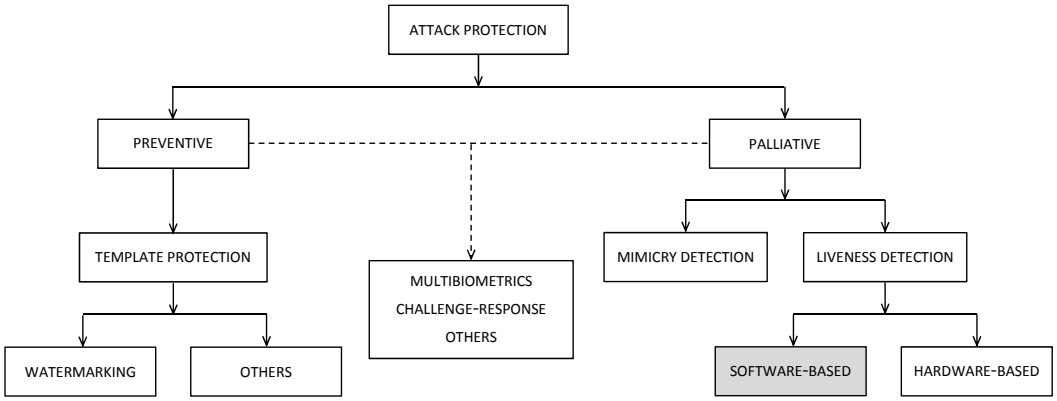


Figure 3.3: Classification of attack protection methods (adapted from [77] for voice modality). Software-based liveness detection techniques, which are analysed in the experimental part of the thesis, appear in coloured box.

For modalities such as face, iris or fingerprint, liveness detection approaches are related to the use of some physiological measure to distinguish between real and fake biometric samples [77]. For a behavioural biometric such as voice, liveness detection refers to the recognition of genuine speech from fake speech i.e. such as that produced or altered by a machine. Countermeasures that detect threats where no electronic manipulation or device is involved in the attack (i.e. impersonation) are refereed as **mimicry detection**.

Liveness detection algorithms can be divided into software-based and hardware-based techniques. In **software-based techniques** the traits are detected once the sample has been acquired by an standard microphone, and are the kind of countermeasures addressed in this thesis. **Hardware-based techniques**, on the other hand, generally capture more or enhanced information beyond that captured normally e.g. liveness detection by designing special fingerprint sensor that also captures blood pressure. While it is not the focus of this thesis, we leave open the possibility of designing special terminals or microphones to protect against spoofing attacks.

The classification presented above is not closed and certain countermeasures, depending on the architecture of the application, can be included in either groups i.e. preventive or palliative. This is the case, for instance, for multi-biometrics or challenge-response countermeasures, among others.

3.2 Previous work

This section reviews past work to evaluate vulnerabilities and to develop spoofing countermeasures. We consider impersonation, replay, speech synthesis and voice conversion. Section 3.2 is based on the work previously published in [67] being some of the contents in which the author has contributed adapted or reproduced in this thesis. A detailed review can be found also in [199]. This survey does not include the contributions in this thesis.

3.2.1 Impersonation

Impersonation refers to spoofing attacks with human-altered voices and is one of the most obvious forms of spoofing and earliest studied.

3.2.1.1 Spoofing

At first glance seems to be that impersonation does not present a risk to ASV systems, due to the fact that usually an impersonator focuses on prosodic and other stylistic features rather than vocal tract features upon which all of today's state-of-the-art voice recognition systems are based. The work in [68, 65] supports this intuitive idea, the former work shows that an imitator can recreate the prosodic characteristics of a target while the latter works the case of an imitator which fails in matching the formant frequencies toward the target, although the opposite is reported in [107].

Also results on direct evaluation of impersonators over ASV systems are inconclusive. The work from *Lau et al.* [115, 114], carried out over the YOHO corpus, shows that impersonators can succeed in overcome a GMM-UBM model. In the first work the impersonators were "close" speakers in the database (according to the scores provided by an ASV system) while in the latter the experiments were reported which six professional impersonators. On the other hand, experiments reported in [127] suggest that even while professional imitators are better impersonators than average people, even they are unable to spoof an ASV system.

Finally, is worth to mention the work in [35, 61]. The authors have observed that some impostor speakers have natural potential to be confused with other speakers. Similarly, certain target speakers may be more easily impersonated than other targets. In all cases, so-called *wolves* and *lambs* leave systems vulnerable to spoofing through the careful selection of target identities. From

the work reported in the literature is still unclear to determine whether impersonation is a real threat to ASV systems.

3.2.1.2 Countermeasures

While the threat of impersonation is not fully understood it is perhaps not surprising that there is no prior work to investigate countermeasures against impersonation. If the threat is proven to be genuine, then the design of appropriate countermeasures might be challenging. Unlike the spoofing attacks discussed below, all of which can be assumed to leave traces of the physical properties of the recording and playback devices, or signal processing artefacts from synthesis or conversion systems, impersonators are live human beings who produce entirely natural speech.

3.2.2 Replay

Replay attacks involve the presentation of speech samples captured from a genuine client in the form of continuous speech recordings, or samples resulting from the concatenation of shorter segments, commonly referred as a 'cut and paste'. Furthermore, replay is a low-technology attack within the grasp of any potential attacker even without knowledge in speech processing. The availability of inexpensive, high quality recording devices can mean that replay is both effective and difficult to detect.

3.2.2.1 Spoofing

In contrast to research involving speech synthesis and voice conversion spoofing attacks where large datasets are generally used for assessment, e.g. NIST datasets, all the past work to assess vulnerabilities to replay attacks relates to small, often purpose-collected datasets typically involving no more than 15 speakers; the lack of appropriate larger-scale datasets mean there is no alternative. While results generated with such small datasets have low statistical significance, differences between baseline performance and that under spoofing are not negligible.

The vulnerability of ASV systems to replay attacks was first investigated in a text-dependent scenario [119] where the concatenation of recorded digits were tested against a hidden Markov model (HMM) based ASV system. Results

showed an increase in the FAR (EER threshold) from 1 to 89% for male speakers and from 5 to 100% for female speakers.

The work in [192] investigated text-independent ASV vulnerabilities through the replaying of far-field recorded speech in a mobile telephony scenario where signals were transmitted by analogue and digital telephone channels. Using a baseline ASV system based on JFA, their work showed an increase in the EER of 1 to almost 70% when imposter accesses were replaced by replayed spoof attacks. A physical access scenario was considered in [195]. While the baseline performance of their GMM-UBM ASV system was note reported, experiments showed that replay attacks provoked an FAR of 93%.

3.2.2.2 Countermeasures

A countermeasure for replay attack detection in the case of text-dependent ASV was reported in [175]. The approach is based upon the comparison of new access samples with stored instances of past accesses. New accesses which are deemed too similar to previous access attempts are identified as replay attacks. A large number of different experiments all relating to a telephony scenario showed that the countermeasures succeeds in lowering the EER in most of the experiments performed.

While some form of text-dependent or challenge-response countermeasure is usually used to prevent replay-attacks, text-independent solution have also been investigated. The same authors in [192] showed that it is possible to detect replay attacks by measuring the channel differences caused by far-field recording [193]. While they show spoof detection error rates of less than 10% it is feasible that today's state-of-the-art approaches to channel compensation will render some ASV systems still vulnerable.

Two different replay attack countermeasures are compared in [195]. Both are based on the detection of differences in channel characteristics expected between licit and spoofed access attempts. Replay attacks incur channel noise from both the recording device and the loudspeaker used for replay and thus the detection of channel effects beyond those introduced by the recording device of the ASV system thus serves as an indicator of replay. The performance of a baseline GMM-UBM system with an EER 40% under spoofing attack falls to 29% with the first countermeasure and a more respectable EER of 10% with the second countermeasure.

3.2.3 Speech synthesis

Speech synthesis, commonly referred to as text-to-speech (TTS), refers to the generation of intelligible speech for any arbitrary text. Applications including this technology can be found nowadays in in-car navigation systems, communication aids for the speech impaired, singing speech synthesizers and and speech-to-speech translation systems, among others.

Speech synthesis technologies can be divided mainly in formant synthesis and concatenative synthesis. Concatenative systems relate to the generation of synthesized speech by concatenating pieces of recorded speech that are stored in a database.

First approaches used a small database of phoneme units called 'diphones' (the second half of one phone plus the first half of the following) while state-of-the-art approaches, referred to as 'unit selection', uses a large database composed by a number of speech units that match both phonemes and other linguistic contexts such as lexical stress and pitch accent to obtain a high-quality natural sounding synthetic speech.

Formant synthesis, on the other hand, generates a speech waveform by a simpler set of rules formulated in the acoustic domain and thus does not use human speech samples at runtime. They are suitable for applications where memory and microprocessor power are especially limited and speech naturalness is not a requirement.

A fourth approach, referred as statistical parametric speech synthesis based on hidden Markov models (HMM) have advanced rapidly over the last decade [208, 120, 27, 210]. One of the primary strengths of the parametric speech synthesis over the traditional unit selection approach is that it requires only a small amount of training data to adapt speaker independent models to a target speaker. It thus allows to build high quality voice models with only a few minutes of adaptation data. Even a few seconds of data is usually sufficient to capture the prominent speaker traits. Hence such a speech synthesis framework becomes an effective tool to carry out spoofing attacks on ASV systems.

3.2.3.1 Spoofing

There is a considerable volume of research in the literature which has demonstrated the vulnerability of ASV to synthetic voices generated with a variety of approaches to speech synthesis. Experiments using formant, diphone, and unit-selection based synthetic speech in addition to the simple cut-and-paste

of speech waveforms have been reported [119, 74, 192].

ASV vulnerabilities to HMM-based synthetic speech were first demonstrated over a decade ago [130] using an HMM-based, text-prompted ASV system [136] and an HMM-based synthesizer where acoustic models were adapted to specific human speakers [131, 132]. The ASV system scored feature vectors against speaker and background models composed of concatenated phoneme models. When tested with human speech the ASV system achieved an FAR of 0% and an FRR of 7%. When subjected to spoofing attacks with synthetic speech, the FAR increased to over 70%, however this work involved only 20 speakers.

Larger scale experiments using the Wall Street Journal corpus containing in the order of 300 speakers and two different ASV systems (GMM-UBM and SVM using Gaussian supervectors) was reported in [52]. Using a state-of-the-art HMM-based speech synthesiser, the FAR was shown to rise to 86% and 81% for the GMM-UBM and SVM systems, respectively. Spoofing experiments using HMM-based synthetic speech against a forensics speaker verification tool BATVOX was also reported in [80] with similar findings. Today's state-of-the-art speech synthesizers thus present a genuine threat to ASV.

3.2.3.2 Countermeasures

Only a small number of attempts to discriminate synthetic speech from natural speech have been investigated and there is currently no general solution which is independent from specific speech synthesis methods. Previous work has demonstrated the successful detection of synthetic speech based on prior knowledge of the acoustic differences of specific speech synthesizers, such as the dynamic ranges of spectral parameters at the utterance level [173] and variance of higher order parts of mel-cepstral coefficients [45].

There are some attempts which focus on acoustic differences between vocoders and natural speech. Since the human auditory system is known to be relatively insensitive to phase [157], vocoders are typically based on a minimum-phase vocal tract model. This simplification leads to differences in the phase spectra between human and synthetic speech, differences which can be utilised for discrimination [55, 198].

Based on the difficulty in reliable prosody modelling in both unit selection and statistical parametric speech synthesis, other approaches to synthetic speech detection use F0 statistics [145, 56]. F0 patterns generated for the statistical parametric speech synthesis approach tend to be over-smoothed and the unit

selection approach frequently exhibits 'F0 jumps' at concatenation points of speech units.

3.2.4 Voice conversion

Several approaches to voice conversion were proposed in the 1980s and 1990s, e.g. [1, 181], and quickly spurred interests to assess the threat to automatic speaker verification (ASV), e.g. [152]. Voice conversion aims to convert or transform the voice of a source speaker (Y) towards that of a specific, target speaker (X) according to a conversion function $\mathcal{F}$ with conversion parameters $\vec{\theta}$:

$$X = \mathcal{F}(Y, \theta) \tag{3.1}$$

The general process is illustrated in Figure 3.4. Most state-of-the-art ASV systems operate on estimates of the short-term spectral envelope. Accordingly, conversion parameters θ are generally optimised at the feature level in order to maximise the potential for spoofing an ASV system which utilises the same or similar feature parameterisations.

While there is a plethora of different approaches to voice conversion in the literature, relatively few have been explored in the context of spoofing. In the following we overview the most common or influential among them.

3.2.4.1 Joint density Gaussian mixture models

As with most voice conversion approaches, and as illustrated in Figure 3.4, the popular *joint density Gaussian mixture model* (JD-GMM) algorithm [99] learns a conversion function using training data with a parallel corpus of frame-aligned pairs $\{(y_t, x_t)\}$. Frame alignment is usually achieved using dynamic time warping (DTW) on *parallel* source-target training utterances with identical text content. The combination of source and target vectors $z = \left[y^{\mathrm{T}} x^{\mathrm{T}}\right]^{\mathrm{T}}$ is therefore used to estimate GMM parameters (component weights, mean vectors and covariance matrices) for the joint probability density of Y and X. The parameters of the JD-GMM are estimated using the classical expectation maximization (EM) algorithm in a maximum likelihood (ML) sense.

During the conversion phase, for each source speech feature vector y, the joint density model is adopted to formulate a transformation function to predict

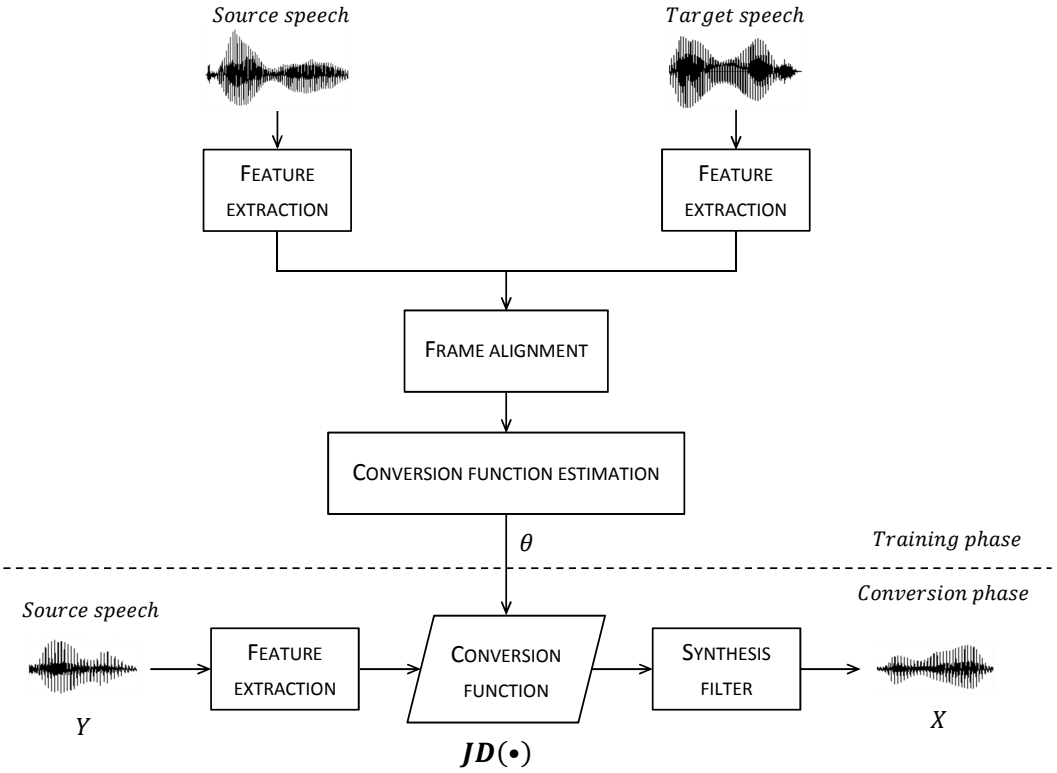


Figure 3.4: An illustration of general voice conversion using, e.g. joint density Gaussian mixture models (JD-GMMs). Figure adapted from [201].

the feature vector of the target speaker according to:

$$\mathcal{JD}(y) = \sum_{l=1}^{L} p_l(y) \left(\mu_l^{(x)} + \Sigma_l^{(yx)} \left(\Sigma_l^{(yy)} \right)^{-1} \left(y - \mu_l^{(y)} \right) \right) \tag{3.2}$$

where $p_l(y)$ is the posterior probability of the source vector y belonging to the l^{th} Gaussian. The trained conversion function is then applied to new source utterances of arbitrary text content at run-time. In addition to parametric voice conversion techniques, *unit selection* – a technique which directly utilizes target speaker segments – is also effective in spoofing ASV [201].

3.2.4.2 Gaussian dependent filtering

The work in [134] extends the concept of JD-GMM to utilise an explicit model of the target speaker at the core of the conversion process. It tests the vulnerabilities of ASV when the vocal tract information in the speech signal of a

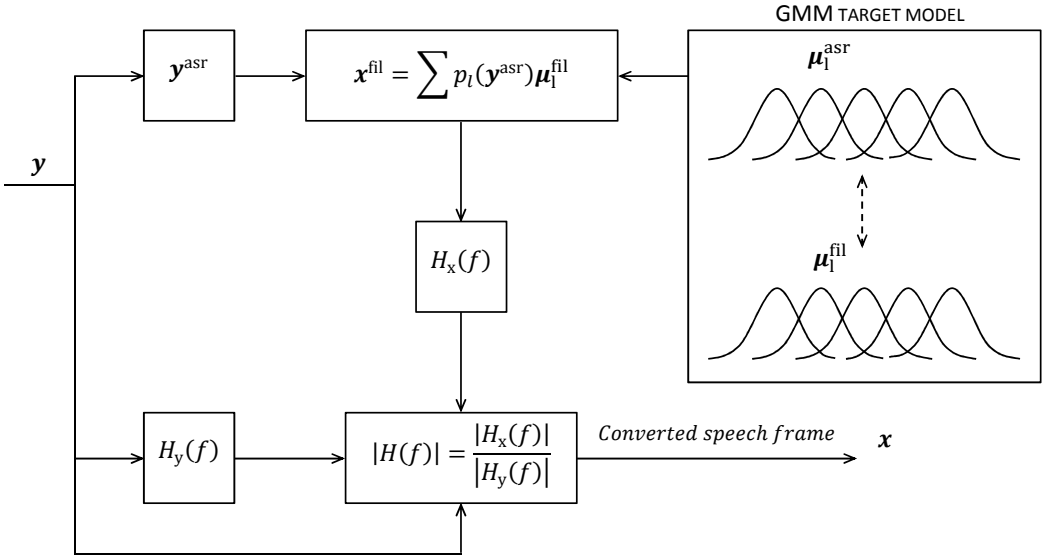


Figure 3.5: An illustration of Gaussian dependent filtering. Figure adapted with permission from [134].

spoofer is converted towards that of the target speaker according to a Gaussian dependent filtering approach. As illustrated in Figure 3.5, the speech signal of a source speaker or spoofer, represented at the short-time frame level and in the spectral domain by $Y(f)$, is filtered as follows:

$$\mathcal{GD}\left(Y(f)\right) = \frac{|H_{\text{x}}(f)|}{|H_{\text{y}}(f)|}Y(f) \tag{3.3}$$

where $H_{\text{x}}(f)$ and $H_{\text{y}}(f)$ are the vocal tract transfer functions of the target speaker and the spoofer respectively and $\mathcal{GD}\left(Y(f)\right)$ denotes the result after voice conversion. As such, each frame of the spoofer's speech signal is mapped or converted towards the target speaker in a spectral envelope sense.

The transfer functions above are estimated according to:

$$H_{\text{x}}(f) = \frac{G_{\text{x}}}{A_{\text{x}}(f)}, \text{ and} \tag{3.4}$$

$$H_{\text{y}}(f) = \frac{G_{\text{y}}}{A_{\text{y}}(f)} \tag{3.5}$$

where $A_{\text{x}}(f)$ and $A_{\text{y}}(f)$ are the Fourier transforms of the corresponding prediction coefficients and G_{x} and G_{y} are the gains of the corresponding residual

signals.

While $H_{\mathrm{y}}(f)$ is obtained directly from Y, $H_{\mathrm{x}}(f)$ is determined from a set of two GMMs. The first, denoted as the automatic speaker recognition (asr) model in the original work, is related to ASV feature space and utilized for the calculation of a posteriori probabilities whereas the second, denoted as the filtering (fil) model, is a tied model of linear predictive cepstral coding (LPCC) coefficients from which $H_{\mathrm{x}}(f)$ is derived. LPCC filter parameters are estimated according to:

$$x^{\mathrm{fil}} = \sum_{l=1}^{L} p_l \left(y^{\mathrm{asr}}\right) \mu_l^{\mathrm{fil}} \tag{3.6}$$

where $p\left(y^{\mathrm{asr}}\right)$ is the posterior probability of the vector y^{asr} belonging to the l^{th} Gaussian in the asr model and μ_l^{fil} is the mean of l^{th} Gaussian belonging to the fil model, which is tied to the l^{th} Gaussian in the asr model. $H_{\mathrm{x}}(f)$ is estimated from x^{fil} using a LPCC to linear prediction (LP) coefficient conversion and a time-domain signal is synthesized from converted frames with a standard overlap-add technique. Resulting speech signals retain the prosodic aspects of the original speaker (spoofer) but reflect the spectral-envelope characteristics of the target while not exhibiting any perceivable artifacts indicative of manipulation. Full details can be found in [134].

3.2.4.3 Spoofing

In the following we review some of the past work which has investigated ASV vulnerabilities to the specific approaches to voice conversion described above.

Even when trained using a non-parallel technique and telephony data, the baseline JD-GMM approach has been shown to increase significantly the false acceptance rate (FAR) of state-of-the-art ASV systems [200]. Even if speech so-treated can be detected by human listeners, experiments involving five different ASV systems showed universal susceptibility to spoofing.

With a decision threshold set to the equal error rate (EER) operating point, the FAR of a joint factor analysis (JFA) system was shown to increase from 3% to over 17% whereas that of an i-vector probabilistic linear discriminant analysis (PLDA) system increases from 3% to 19%. The unit-selection approach was shown to be even more effective and increased the FARs to 33% and 41% for the JFA and PLDA systems respectively.

The work reported in [134] investigated vulnerabilities to voice conversion through the Gaussian dependent filtering of the spectral-envelope. Voice conversion was applied using the same feature parameterisations and classifier as the ASV system under attack. Results thus reflect the worst case scenario where an attacker has full knowledge of the recognition system and show that the EER of a GMM-based ASV system increases from 10% to over 60% when all impostor test samples were replaced with converted voice.

3.2.4.4 Countermeasures

As the above shows, current ASV systems are essentially 'deaf' to conversion artifacts caused by imperfect signal analysis-synthesis models or poorly trained conversion functions. Tackling such weaknesses provides one obvious strategy to implement spoofing countermeasures.

Some of the first work to detect converted voice [198] draws on related work in synthetic speech detection and considers phase-based countermeasures to JD-GMM and unit-selection approaches to voice conversion. The work investigated two different countermeasures, referred to as the cosine normalization and frequency derivative of the phase spectrum. Both countermeasures aim to detect the absence of natural speech phase, an artifact indicative of converted voice.

The two countermeasures are effective in detecting converted voice with EERs as low as 6.0% and 2.4% respectively. In [200], the detector is combined with speaker verification systems for anti-spoofing. With a decision threshold set to the equal error rate (EER) operating point, baseline FARs of 3.1% and 2.9% for JFA and PLDA systems respectively fall to 0% for JD-GMM voice conversion attacks and to 1.6% and 1.7% for unit-selection attacks.

Phase-based countermeasures may be bypassed, however, by approaches to voice conversion which retain natural speech phase, i.e. approaches such as Gaussian-dependent filtering [134]. This problem is addressed in this thesis.

For further details on spoofing and countermeasures for voice conversion readers are referred to [141].

3.2.5 Summary

A comparative study of vulnerabilities and countermeasures is summarized in Table 3.1 for text-independent ASV systems and the four attacks described

before. The comparison is reported in terms of two factors, denoted *effort*[1] i.e. the specific skill, expertise, information or equipment to perform the attack and *risk*, which reflects the effectiveness in each approach in provoking higher false acceptance rates.

Attack	Effort	Risk	Countermeasure
Impersonation	High	Low	Non-existent
Replay	Low	High	Low
Speech synthesis	High	High	Medium
Voice conversion	High	High	Medium

Table 3.1: A summary of the required effort and risk of the four spoofing attack approaches, and the availability of countermeasures for automatic speaker verification.

The literature review presented in Section 3.2 as well as Table 3.1 are adapted from the work in [67, 199]. This work also acknowledge the challenging task of a fair comparison of the results presented in the literature due to the multitude of extremely different experimental conditions. The reading and interpretation of the table and the subsequent should be thus taken with care.

High-effort attacks require either specific skills, e.g. impersonation, or high-level technology neither available for the mass nor easy to use[2], e.g. speech synthesis and voice conversion. On the other hand, a low-effort attacks such as replay attacks can be performed without any specific expertise nor any sophisticated equipment.

The practical risk of an ASV system to be fooled by impersonation seems to depend on the skill of the impersonator, the similarity of the attacker's voice to that of the target speaker, and on the recognizer itself. On the contrary, replay attacks, voice conversion and voice synthesis remain highly effective in all previous studies.

Finally, it can be observed that currently there are no impersonation countermeasures, which seems logical since the risk that impersonation present to ASVs is still not fully understood and its detection is troublesome. Surprisingly, while countermeasures for speech synthesis and voice conversion have attracted most of the attention of the scientific community, there are only few publications on countermeasures against replay attacks, despite of the fact of the low-effort to perform and high-risk they present to ASV systems.

[1] In this thesis we adopt the terms *effort* and *risk* instead of the terms *accessibility* and *effectiveness* used in [67, 199], respectively

[2] There is evidence that suggest that in the future both conditions will change

3.3 Datasets, protocols & metrics

A general approach to assess spoofing vulnerabilities and the performance of anti-spoofing countermeasures consist in two steps. First, a speech database is used to evaluate baseline ASV performance. These experiments assess both genuine client and naive impostor trials. Second, the naive impostor trials are replaced with spoofed trials and the experiment is repeated. The aim is then to evaluate the degradation in performance, perhaps in terms of the equal error rate (EER) or false acceptance rate (FAR), usually derived from detection error trade-off (DET) profiles [66, 67].

The performance of anti-spoofing countermeasures is typically assessed in isolation from ASV, using the same speech database of genuine and spoofed trials used to assess vulnerabilities. Performance can again be assessed in terms of the EER or FAR. Some researchers have also investigated the resulting effect of countermeasures on ASV performance, e.g. [5]. Results furthermore reflect the performance of non-standard ASV systems.

The lack of consensus in the biometry community related to spoofing and countermeasure evaluations can be reflected for instance in the number of existing notations (synonyms) presented in the literature [48]. From the bulk of different approaches, the most accepted is to define a third error to describe the ratio of incorrectly accepted spoofing attempts [98], denoted spoofing FAR (SFAR).

SFAR and FAR are synonyms, but with the difference that defining a third error (together with FRR and FAR) formulates the spoofing evaluation as a three class problem. Spoofing evaluations are thus reported by comparison of values of FAR and SFAR.

A thorough review of the spoofing and countermeasures evaluations used in the literature for different modalities can be found in [48]. Also this work proposes a new framework for spoofing and countermeasure evaluation denoted Expected Performance and Spoofability (EPS) framework. Although it is worth mentioning this novel and promising work, it was not available earlier in this thesis and is thus not used in this work.

Despite of the fact of the notable exception mentioned above, due to the novelty of such work there are currently no standard large-scale datasets, protocols or metrics which might otherwise be used to conduct evaluations with a fairer sense and more comparable results. There is a thus a need to define such standards in the future.

3.3.1 TABULA RASA evaluations

Candidate standards were drafted within the scope of the EU FP7 TABULA RASA project[3]. Here, independent countermeasures preceding biometric verification are optimized at three different operating points where thresholds are set to obtain FARs (the probability of labeling a genuine access as a spoofing attack) of either 1, 5 or 10%. Samples labeled as genuine accesses are then passed to the verification system[4].

Performance is assessed using four different DET profiles[5], examples of which are illustrated in Figure 3.6. The four profiles illustrate performance of the baseline system with naive impostors, the baseline system with active countermeasures, the baseline system where all impostor accesses are replaced with spoofing attacks and, finally, the baseline system with spoofing attacks and active countermeasures.

Consideration of all four profiles is needed to gauge the impact of countermeasure performance on licit transactions (any deterioration in false rejection - difference between 1st and 2nd profiles) and improved robustness to spoofing (improvements in false acceptance - difference between 3rd and 4th profiles). While the interpretation of such profiles is trivial, different plots are obtained for each countermeasure operating point.

In this evaluation the countermeasure evaluated stand-alone is subjected to two kind of errors denoted False Living Rate (FLR), which represents the percentage of fake data misclassified as real, and False Fake Rate (FFR), which computes the percentage of real data assigned to the fake class. They are similar to FAR and FRR, respectively. The lower these two errors, the better the performance of the countermeasure. The point at which FLR=FFR is called Average Classification Error (ACE) of the fake detection task, which is similar to the Equal Error Rate (EER).

[3] `http://www.tabularasa-euproject.org/`

[4] In practice samples labeled as spoofing attacks cannot be fully discarded since so doing would unduly influence false reject and false acceptance rates calculated as a percentage of all accesses.

[5] Produced with the TABULA RASA Score toolkit `http://publications.idiap.ch/downloads/reports/2012/AnjosIdiap-Com-02-2012.pdf`

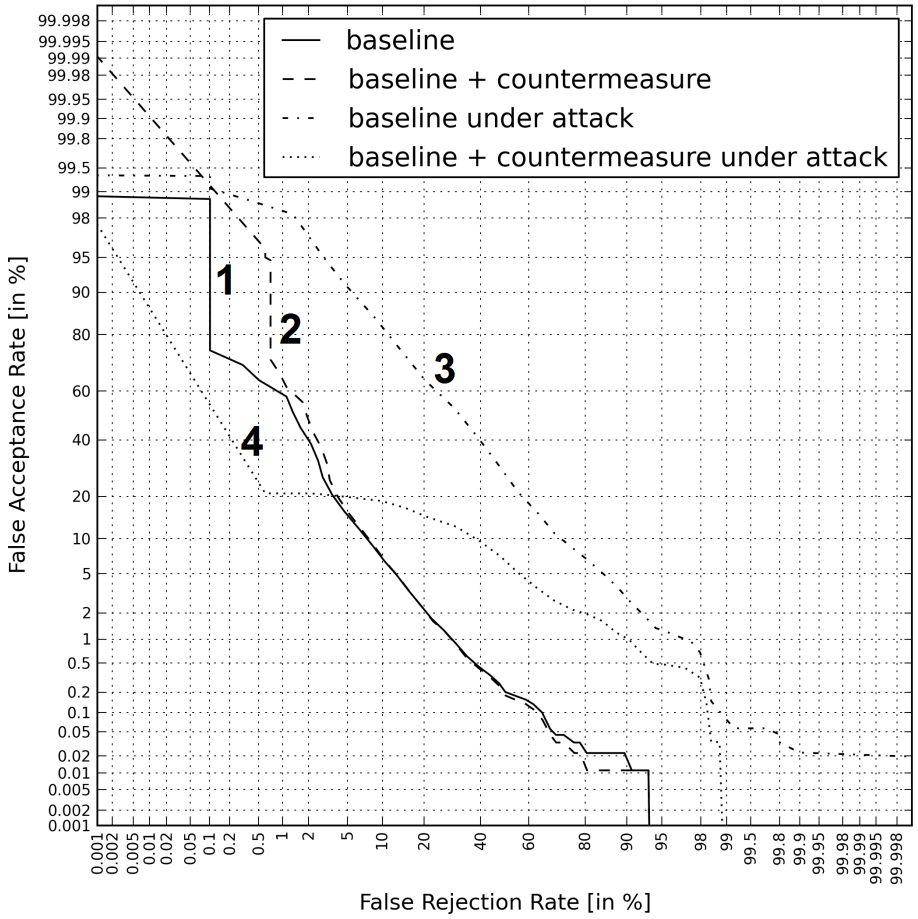


Figure 3.6: An example of four DET profiles needed to analise vulnerabilities to spoofing and countermeasure performance, both on licit and spoofed access attempts. Results correspond to spoofing using synthetic speech and a standard GMM-UBM classifier assessed on the male subjet of the NIST'06 SRE dataset.

3.3.2 Spoofing datasets

While some work has shown the potential for detecting spoofing without prior knowledge or training data indicative of a specific attack [5, 198, 11], all previous work is based on some implicit prior knowledge, i.e. the nature of the spoofing attack and/or the targeted ASV system is known. While training and evaluation data with known spoofing attacks might be useful to develop and optimise appropriate countermeasures, the precise nature of spoofing attacks

can never be known in practice. Estimates of countermeasure performance so obtained should thus be considered at best optimistic.

Furthermore, some of the past work was also conducted under matched conditions, i.e. data used to learn target models and that used to effect spoofing were collected in the same or similar acoustic environment and over the same or similar channel. The performance of spoofing countermeasures when subjected to realistic session variability is then unknown.

3.4 Discussion

In general, most of the studies reported in the literature assess ASV vulnerabilities under specific conditions i.e. one scenario, one spoofing attack and usually one ASV system, being the attack among impersonation, replay, speech synthesis and voice conversion. The current literature supports the idea about a common believe that the attacks are limited to the four previously mentioned, while deeper, more comprehensive efforts to understand the mechanisms behind the effectiveness of spoofing attacks are still non-existent.

Furthermore, a problem that not directly linked to evaluation methodologies relates to the fact that almost all ASV spoofing countermeasures proposed thus far are dependent on training examples indicative of a specific attack. Given that the nature of spoofing attacks can never to known in practice, and with the variety in spoofing attacks particularly high in ASV, there is a need to investigate new countermeasures which generalise well to unforeseen attacks.

These and other issues are addressed in this thesis. For a comprehensive literature review of the topic, readers are referred to [67, 199].

Part II

CONTRIBUTIONS

Chapter 4

Spoofing assessment

It is widely acknowledged that automatic speaker verification (ASV) systems are vulnerable to spoofing. A growing body of independent work has now demonstrated the vulnerability of ASV systems to spoofing through impersonation, replay attacks, voice conversion and speech synthesis. An overview of these attacks is introduced in Section 3.2

This chapter reports our assessment of ASV system vulnerabilities and spoofing attacks. It aims to bring some insight to the vulnerabilities in ASV systems through the study of the mechanisms behind spoofing attempts. In particular, this chapter aims to identify some of the factors that limit the current knowledge on spoofing attacks.

The remainder of the chapter is as follows. Section 4.1 presents an initial analysis of potential vulnerabilities in ASV systems with respect to spoofing attacks. Section 4.2 is divided in two part. First, it investigates the feasibility of new threats to ASV systems beyond the types already known. In particular, an approach to artificial signals is presented to highlight the vulnerability of ASV systems to entirely artificial, non-speech-like tone signals. Second, it accesses spoofing attacks in terms of the effort. Finally, Section 4.3 discusses about common issues related to the vulnerability evaluation of ASV systems.

4.1 Vulnerabilities in ASV systems

In this thesis we are interested in the analysis and detection of successful spoofing attacks[1]. The first logical step is to define the meaning of *successful* and the meaning of *spoofing attack* followed in this thesis.

In the context of this thesis, *successful* means "able to fool an ASV system". The modules of the generic ASV system illustrated in Figure 2.1(b) can be regrouped so that the ASV can be seen as the cascade of speaker detection and speaker recognition classifiers (Figure 4.1). To be a successful spoofing

[1] despite of argument that an unsuccessful attack is not spoofing

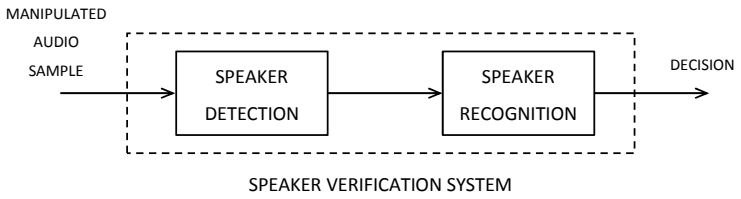


Figure 4.1: Simplified version of the ASV architecture illustrated in Figure 2.1(c) in which the ASV modules are regrouped in speech detection and speech recognition. A successful attack must overcome both modules.

attack is thus *sufficient* for the input signal to overcome the speaker detection module **and** the speaker recognition classifiers of an ASV system.

The previous reasoning highlights the importance of the analysis of vulnerabilities in ASV systems. In this sense, we observe that all the modules conforming the generic ASV system introduced in Section 2.2 present weak points that can be exploited by a spoofer.

All feature representations including short-term spectral features, prosodic features as well as high-level features introduced in Section 2.2.2, are potentially vulnerable to spoofing attacks. State-of-the-art voice conversion systems and speech synthesizers can generate speech signals whose vocal tract characteristics and/or prosodics reflect those of a targeted speaker. Also prosodic characteristics and language content and speaker behavior related to high-level features may also be mimicked through impersonation, while all the acoustic characteristics previously mentioned are intrinsically contained in the prerecorded speech samples used for replay attacks.

SAD systems also present weaknesses. While model and phoneme-based SAD might provide greater robustness to spoofing, energy-based approach might be preferred in practice (Section 2.2.1). Energy-based SAD is, however, the most vulnerable to spoofing, being easily overcome with any high-energy input signal. This, in addition to the fact that almost all approaches to speech detection work at the frame level, open the possibility of the ASV systems to be vulnerable to a wide family of new threats.

Common characteristic shared by all the biometric systems is that their models are learned by utilizing entirely biometric samples. For speaker recognition, usually formulated as a statistical hypothesis testing problem, both the null and the alternative hypothesis are modeled with speech samples. This characteristic could be arguably a vulnerability in the face of spoofing. For instance, the behavior of the recognition module against an unseen sample e.g. a non-

speech spoofing signal can be unpredictable. This discussion is extended in Section 4.2.

Moreover, most approaches to speaker modelling have their roots in the standard GMM and thus they model the feature distributions and disregards temporal sequence information. Hence, typical speech synthesis and voice conversion algorithms which assume independent frames are effective as spoofing attacks.

4.2 Spoofing attacks

As mentioned before, the literature on spoofing is limited to four different attacks, including classical spoofing attacks such as impersonation or replay as well as more advanced attacks such as voice conversion and speech synthesis, all shown to provoke significant increases in the false acceptance rate of state-of-the-art ASV systems. Here the common approach is to generate spoofing samples by matching/imitating the acoustic features of the targeted speaker.

The sufficient condition used in this thesis to define successful spoofing attacks includes but is not limited to the attacks mentioned above. In this sense, from the analysis in Section 4.1 we note that the speech detection and the recognition module module also present their own weaknesses. This fact suggest that acoustic features matching may not be necessarily the only approach to spoofing. In particular, this section addresses spoofing with non-speech signals.

Common to all the previous work is the assumption that spoofing attacks are performed with speech of reasonable quality. It is assumed either that non-speech signals are rejected by speech activity detection (Section 2.2.1) or that the speaker recognition classifier is inherently robust to non-speech signals.

The first option is not necessarily true in real applications (e.g. energy-based SAD might be overcome by high-energy non-speech signals) and the latter is unfounded, since the speaker recognition techniques are not designed to cope with such signals. In this case the output is unpredictable.

This section aims to identify this and some other factors that limit the current knowledge on spoofing attacks. In particular, Section 4.2.1 present an approach to artificial signals which is later shown to provoke significant increases in false acceptances of typical ASV systems.

To the best of our knowledge, this work, also published in [13], is the first

to consider the potential vulnerability of ASV to non-speech signals. Section 4.2.2 revisits ASV vulnerabilities in terms of practical risk. Finally, Section 4.2.3 aims to provide further insight on the mechanisms behind spoofing attacks.

4.2.1 Attacks through artificial signals

In this section we propose a procedure to design an artificial signal capable to provoke high scores in typical ASV system. The contents in the following section are reproduced from the author's own work previously published in [13, 12].

Our approach to test the vulnerabilities of ASV systems to spoofing combines voice conversion and the notion of so-called replay attacks, where a genuine-client recording is replayed to a biometric sensor, here a microphone. Particularly if it is equipped with channel compensation routines, then it is entirely possible that an ASV system may be overcome through the replaying of a client speech signal X; this is a conventional replay attack.

However, certain short intervals of contiguous frames in X, e.g. those corresponding to voiced regions, will give rise to higher scores or likelihoods than others. The probability of a replay attack overcoming an ASV system can thus be increased by selecting from X only those components or frames which provoke the highest scores. The resulting signal will not sound anything like intelligible speech but this is of no consequence if we assume, as is generally the case, that the ASV system in question uses only energy and/or pitch-based speech activity detection (SAD) and does not incorporate any form of speech quality assessment.

Here we consider an attack based upon the extraction and replaying of a short interval or sequence of frames in $X = \{x_1, ..., x_m\}$ which gives rise to the highest scores.

Let $T = \{c_1, \ldots, c_n\}$ be such an interval short enough so that all frames in the interval provoke high scores, but long enough so that relevant dynamic information (e.g. delta and acceleration coefficients) can be captured and/or modelled. In order to produce a replay recording of significant duration, T can be replicated and concatenated any number of times to produce an audio signal of arbitrary length. In practice the resulting concatenated signal is an artificial, or tone-like signal which reflects the pitch structure in voiced speech.

Even though such signals can be used themselves to test the vulnerabilities

of ASV systems, their limits can be more thoroughly tested by enhancing the above approach further through voice conversion. Each frame in T can be decomposed and treated in a similar manner as described in Section 3.2.4.2.

The short interval or sequence of frames in T can be represented as:

$$S_T = \{S_{c_1}(f), S_{c_2}(f), ..., S_{c_n}(f)\}, \text{ and} \quad (4.1)$$
$$H_T = \{H_{c_1}(f), H_{c_2}(f), ..., H_{c_n}(f)\} \quad (4.2)$$

Each frame $c_i \in T$ can be reconstructed from their corresponding elements in S_T and H_T. While S_T captures the excitation source, which has little influence on ASV, H_T captures the vocal tract response from which cepstral features are extracted. Since it has no impact on ASV the phase information in (4.2) is discarded in practice.

Therefore, each frame c_t belonging to S is transformed in the frequency domain with voice conversion (Equation (3.3)) where we now have:

$$\mathcal{AS}\left(C(f)\right) = \frac{|F_{\mathrm{c}}(f)|}{|H_{\mathrm{c}}(f)|}C(f) \quad (4.3)$$

The set of excitations $S_{\mathrm{T}} = \{S_{\mathrm{c}_1}(f), S_{\mathrm{c}_2}(f), \ldots, S_{\mathrm{c}_n}(f)\}$ remains the same as the ones extracted from T, then the problem is reduced to identify a set of filters $H^*_{\mathrm{T}} = \{F_{\mathrm{c}_1}(f), F_{\mathrm{c}_2}(f), \ldots, F_{\mathrm{c}_n}(f)\}$.

Therefore, we aim to estimate a new set of transfer functions F_T to replace H_T in Equation (4.2) in order to synthesise a new artificial signal more likely to spoof the ASV system, and consequently a more stringent test of vulnerabilities. In the same way as in Equation (3.5), F_T can be split into gains G_t and frequency responses $A_t(f)$ giving sequences:

$$G_T = \{G_{c_1}, G_{c_2}, ..., G_{c_n}\} \quad (4.4)$$
$$A_T = \{A_{c_1}(f), A_{c_2}(f), ..., A_{c_n}(f)\} \quad (4.5)$$

where each $A_c(f)$ is obtained from p prediction coefficients for frame c, i.e. $P_c = \{a_{ic}\}_{i=1}^{p}$. The prediction coefficients for the sequence are denoted by P_T.

$$P_T = \{P_{c_1}, P_{c_2}, ..., P_{c_n}\} \quad (4.6)$$

We then seek a set of parameters to synthesize a new signal which maximises the ASV score according to the following objective function:

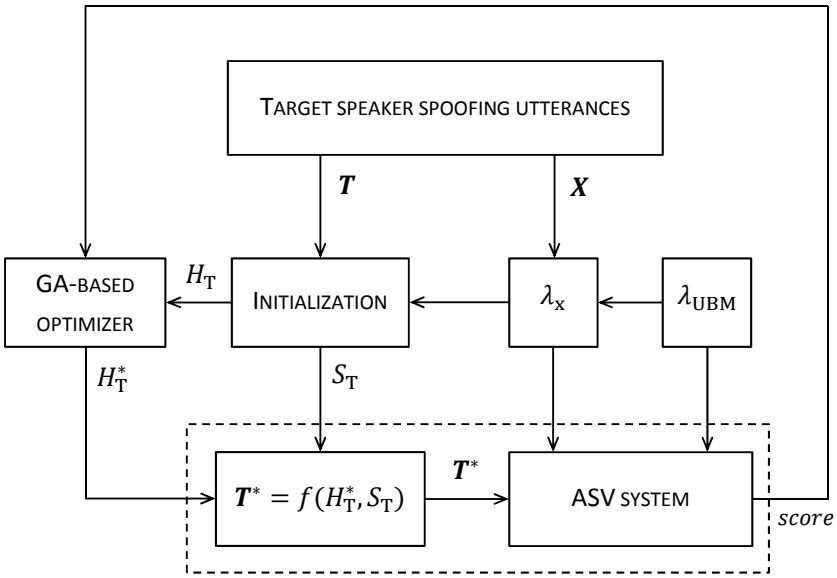


Figure 4.2: Schematic representation of the optimization loop

$$(P_T^*, G_T^*) = \arg\max_{P_T, G_T} l(f(P_T, G_T, S_T), \lambda_X, \lambda_{UBM}) \tag{4.7}$$

where, $f()$ is a function which reconstructs a signal from the parameters G_T, P_T and S_X, and $l()$ is the ASV function that scores the generated signal with respect to the target speaker model λ_X and the universal background model λ_{UBM}. Note that the ASV system has a dual role both in identifying the short interval T and in the subsequent optimisation process.

The $(p+1)*n$ variables comprising the prediction coefficients, gains and ASV score in Equation (4.7) are continuous valued and the optimization problem is non-convex and possibly discontinuous in nature. In our work Equation (4.7) is maximised with a genetic optimisation algorithm. Genetic algorithms are well-suited to the stochastic nature of the speech signals and have been applied previously in related work, e.g. voice conversion [215, 213] and speech synthesis [149].

A schematic representation of the optimization problem is illustrated in Figure 4.2. The target speaker spoofing utterance X is used to learn the target speaker model λ_X as well as in the selection of the short segment T for constructing the artificial signals. The dashed block represents the optimisation objective function. The resulting waveform is illustrated in Figure 4.3.

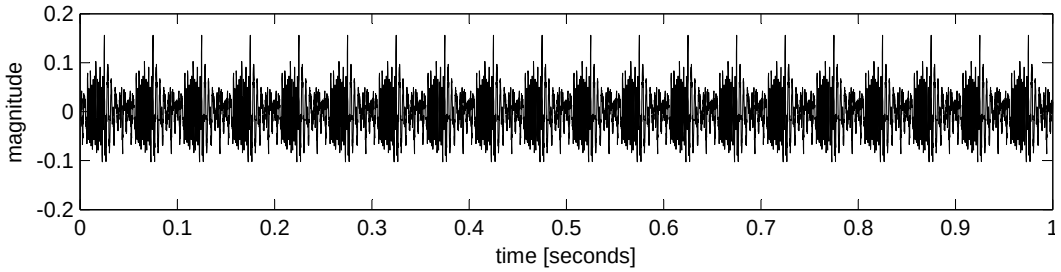


Figure 4.3: Waveform of the resulting tone-like artificial signal.

4.2.2 Attacks in terms of effort

Common to the bulk of previous work is the consideration of attacks which require either specific skills, e.g. impersonation, or relatively high-level technology, e.g. speech synthesis and voice conversion, being replay attacks the only one that can be performed with relative ease. This section is a first attempt to address spoofing in terms of practical risk.

The former necessarily implies the existence of low and medium effort attacks not know yet. Thus, this section aims to determine whether or not other signals, apart from the four types widely known and the one presented in previous section, can also be used to spoof an ASV system.

As illustrated in Figure 4.1, a successful spoofing attack must overcome both speech detection and speaker recognition modules which together constitute a typical approach to automatic speaker verification (ASV). This section discusses ASV spoofing in terms of (i) non-speech signals with the potential to overcome speech detection and (ii) the effort required by the would-be spoofer to implemented the attack.

An exhaustive analysis of new threats is out of the scope of this thesis. However, to motivate further research, we report experimental work in which the speech in the impostor trial utterances are replaced by white noise and tested against a collection of state-of-the-art ASV systems. The results, presented in Section 5.3.2.2, show that when the impostors are replaced with white noise the false acceptances of most of ASV systems increases.

Table 4.1 extends Table 3.1 by including the two novel attacks introduced in this chapter. The second column illustrates non-speech signals whereas the third column illustrates speech signals.

The attacks presented in this table have all been shown in previous work to have the potential to overcome an ASV system with an energy-based SAD.

Effort	Non-speech signals	Speech signals
Zero		Naive impostors
Low	White noise (Section 4.2.2)	Replay attacks [119, 8]
Medium		Replay (cut and paste) [192]
High	Artificial signals [13, 12]	Impersonation [119, 28] Voice conversion [153, 30] Speech synthesis [130, 52]

Table 4.1: A classification of speech and non-speech signals in terms of effort required to spoof an ASV system. The risk of all the mentioned attacks is evaluated in this document, except for replay-related attacks (replay attacks related experiments are reported in [8].

Replay, impersonation, voice conversion and speech synthesis are all natural speech attacks, whereas artificial signals introduced in Section 4.2.1 and also in [13, 12] are entirely non-speech-like, yet still have the potential to spoof an ASV system. Even so, all of these attacks can only be implemented with a certain level of expertise and with target training data.

Table 4.1 also classifies example spoofing signals in terms of the effort required to implement each attack. For example, white noise generation requires low effort and does not require target speech data, which decreases even more the level of effort. Others, such as replayed speech, for example, require a low level of expertise and equipment or medium level of expertise in the case of cut and paste attacks.

Specially-crafted artificial signals and targeted speech synthesis in contrast, require a specific high-level of expertise and, while they are highly effective in overcoming ASV, they are arguably beyond the means of the average would-be spoofer. For comparison purposes, this table includes naive impostors, which "implement" the attack without effort.

4.2.3 Discussion

The work in this section brings some insight into the field of spoofing attacks by expanding the number of known threats. It shows that an ASV systems have the potential to be fooled with non-speech attacks e.g. artificial signals, or low-effort attacks other than replay e.g. white noise.

The approach to artificial signals introduced in Section 4.2.1 basically modifies the spectral envelope of a small number of short-term speech frames in order to synthesize a tone-like audio signal which maximises the score of a given ASV system. Hence, it is expected that the effectiveness of the attack will be dependent on the similarities between the targeted ASV system and the ASV system used to synthesize the attack i.e. technologies used, systems configuration, etcetera. This hypothesis is investigated in the experimental work in this thesis

Our experimental work in the next chapter[2] suggest that an audio signals containing purely white noise can be marginally a better attack than naive impostor, even though the increase in false acceptances is not significant.

Although the risk of attacks with white noise may be negligible with respect to a naive impostor, the former becomes relevant if, for instance, an spoofer has the interest and means to target a group or an unlimited number of speaker. We note that, opposite to all previous attacks, white noise is an attack which overcome speaker recognition without the use of target-specific training data.

Finally, even thought there is no previous reports of such *generalized spoofing attacks*, inferences can be made from observations in the literature. As mentioned in Section 3.2.1 some impostor speakers denoted *wolves* have natural potential to be confused with other speakers. Would be possible to synthesize a generalized spoofing attack by voice conversion, speech synthesis or artificial signal based on a model of such speakers?

4.3 Limitations of current spoofing assessments

This thesis has already pointed out about the growing bulk of work to assess the impact of spoofing on speaker recognition and the performance of anti-spoofing countermeasures. Thus far, the community has concentrated on four predominant forms: impersonation, replay, speech synthesis and voice

[2]We advance some of the experimental results for the sake of clarity

conversion.

Whatever the form of attack, however, there are no standard databases, protocols or metrics which are adequate in their original form for research in spoofing and anti-spoofing. As a result, most studies involve either standard speech databases which are modified according to some particular non-standard spoofing algorithm, or often-small, purpose-collected databases. In neither case are results produced by one study, meaningfully comparable to those produced by another.

The lack of consensus on best practices and techniques to evaluate biometric system vulnerabilities and spoofing detection is a key issue for most of the modalities [16], although seems to be more pronounced for the voice modality if compared with face and fingerprint, for instance. Apart from a notable exception in [74], for the best of our knowledge there is still any serious initiative to define the pillars toward the development of a standardized evaluation framework.

Our work in [7] provides an analysis of the issues related to current vulnerability evaluation of ASV systems. This section, aims to contribute to this previous work by providing an alternative explanation of some of the limitations and weaknesses in the design of spoofing datasets from the point of view of the two key stages involved in an spoofing attempt, denoted biometric data acquisition and spoofing signal insertion [70], respectively.

4.3.1 Acquisition point & insertion point

In general, a spoofing attack requires prior knowledge of the targeted speaker; except for impersonation, this prior knowledge is usually a speech recording. The process to obtain such a prior knowledge, referred as biometric signal acquisition, can be grouped according to their **acquisition point** into three main categories: (1) acquisition at the transmission level, (2) acquisition at the sensor level, and (3) no acquisition.

An speech signal is captured at the **transmission level** when it is stored after microphone and channel transmission, possibly the same microphone and/or channel used during recognition. Usually they consist in 8KHz sampled signals captured in telephony or remote authentication scenarios, being a typical example to intercept and record a phone call of the targeted speaker at some point of the transmission or at the end of the secondary terminal.

Acquisition at the **sensor level** is related to the use of a portable recorder

or "in site" microphone to capture the speech signal. This approach could be advantageous for the spoofer with respect to (1) in the sense that there is not channel distortion and also this approach permits higher quality recordings i.e. speech sampled at 48KHz. On the other hand, it could be affected by reverberation or other form of distortion due of a far field recordings. Finally, there is **no acquisition** at all in the case of impersonation or a (hypothetical) generalized spoofing attack.

Insertion points are closely related with use cases defined in Section 1.3.2. Hence, an attack perpetrated at the sensor level (point 1 in Figure 3.2) is typical of **physical access scenarios**, where the microphone and transmission channel of the recognition system are secured, and requires the use of a device to playback the spoofing signal. On the other hand, attacks at the transmission level (point 2 in Figure 3.2) are possible in telephony or remote authentication scenarios, where neither the sensor nor the transmission channel are secured.

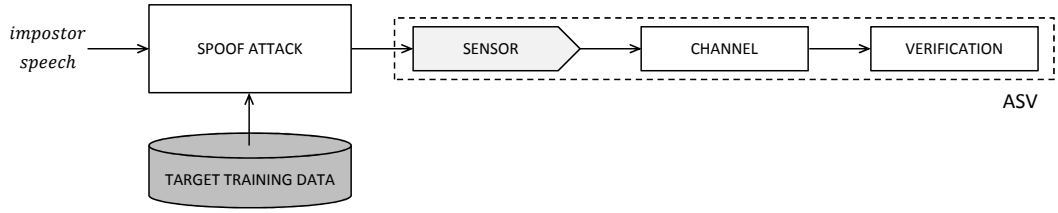
Although in **mobile/telephony scenarios** are also propitious to sensor-level attacks, there is a common believe that avoiding the distortion due of the environment and microphone will increase the changes of the spoofing attacks to succeed. In any case, this thesis assumes attacks in mobile/telephony scenarios to be perpetrated at the transmission level, which is considered the worst case scenario.

4.3.2 Spoofing datasets design

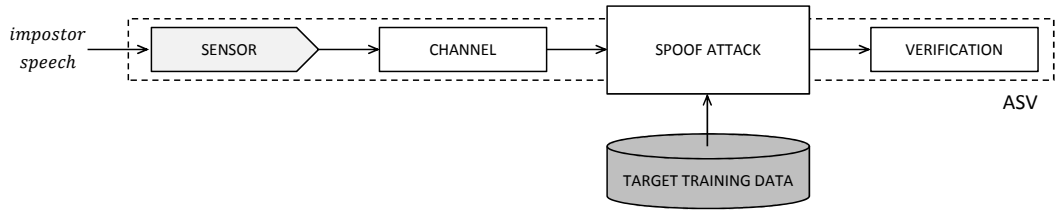
Assumptions on biometric data acquisition and spoofing signal insertion play an important role the collection spoofing databases and design of security evaluations. Most of the contents in this section are reproduced from the author's own work previously published in [7].

We note that standard, large scale databases contain utterances captured only at the transmission level. Hence, even thought impersonation and replay attacks are the least sophisticated and therefore the most accessible attacks [67], the research to develop anti-spoofing systems capable of detecting impersonation and replay is limited by the use of purpose-made, small or medium scale databases; since there are no standard databases of impersonated or replayed speech.

In addition, it is not possible, or at least extremely troublesome to adapt existing, standard databases for such research. For instance, for attacks at



(a) Attack at the sensor level



(b) Simulated attack

Figure 4.4: A comparison of sensor-level spoofing attacks and the general approach to simulate spoofing attacks using standard databases (transmission-level spoofing).

the sensor level such as replay attack, utterances can be either re-recorded following specific protocols (not developed yet) or either the utterances can be artificially convoluted with different impulse responses, although the validity of the latter approach must be investigated. In any case, we do not consider impersonation or replay attacks any further in this thesis.

Speech synthesis [130] and voice conversion [152] attacks have attracted a great deal of attention. Even if they are the least accessible [67] (they involve sophisticated technology), there is evidence that both forms of attack can provoke significant degradation in ASV performance [66]. In addition, research involving speech synthesis and voice conversion attacks can be performed using adapted, standard databases.

We note, however, that while the adaptation of standard datasets to assess spoofing vulnerabilities is the common approach, such a setup only addresses attacks (insertion) at the transmission level and it is not reflective of the traditional consideration of spoofing at the sensor level defined in the previous section.

Figure 4.4 illustrates the difference. As illustrated in Figure 4.4(a), an attacker will normally obtain examples of the target's speech in order to adjust or optimise a spoofing attack at the sensor level. Speech signals are then subjected to acquisition and channel or coding effects before verification. This

Scenario		Evaluation	
insertion point	acquisition point	datasets	attacks
transmission	transmission	large-scale	Voice conversion [153, 30] Speech synthesis [130, 52] Artificial signals [13, 12]
	sensor	small-scale	Voice conversion [153, 30] Speech synthesis [130, 52]
	no-acquisition	-	-
sensor	transmission	-	-
	sensor	small-scale	Replay attack [119, 192]
	no-acquisition	small-scale	Impersonation [119, 28]

Table 4.2: Spoofing databases in the literature in terms of insertion point and acquisition point. We note that (i) there is a lack of work in sensor-level scenarios and (ii) only transmission-transmission scenarios are tested with large-scale databases.

process differs from the practical setup illustrated in Figure 4.4(b). Here the spoofing attack is performed post-sensor, immediately before verification.

If we assume that the sensor, channel and spoofing attack are all linear transforms, then the order in which they occur is of no consequence; the two setups illustrated in Figure 4.4 are equivalent and this setup is also valid to assess sensor-level attacks. This assumption, however, is unlikely in real applications.

Table 4.2 groups relevant work in vulnerability assessment against spoofing, where here the attacks are represented as the combination of insertion-acquisition points (Section 4.3.1). The table shows that only transmission-transmission scenarios are tested with large-scale databases, while transmission-sensor scenarios are tested with in-house collected databases. Furthermore, the (unlikely?) transmission-no-acquisition scenario (e.g. to record an impersonator an insert the signal at the transmission level) is not still considered in the literature.

Experiments with sensor-level setups correspond only to replay and impersonation attacks, while there is still not work related to sensor-transmission scenarios. From this table we conclude that sensor level spoofing attacks is underestimated and warrants wider attention.

In [7] we provide a brief overview of the most significant databases used in prior work in ASV spoofing involving both text-independent databases and recent efforts using text-dependent databases.

Chapter 5

Evaluation: ASV systems & spoofing

This chapter introduces our own analysis to measure the effectiveness of direct attacks to a ASV systems and related issues, in order to provide an insight as to the vulnerability of different recognition systems are to these threats.

This chapter defines the specifications related to ASV performance and vulnerability assessment, including ASV systems and spoofing attacks description and setup, biometric and spoofing databases, and protocols and metrics adopted for each evaluation.

In Section 4.3 we have identified two complementary scenarios for the use cases reported in Chapter 1. They are the physical access and mobile/telephony scenarios respectively. This thesis investigates the latter, while experiments for physical access scenario are out of the scope of this thesis[1].

5.1 Specifications for performance evaluation

The NIST speaker recognition evaluation datasets are the de facto standard, are used in all state-of-the-art research and are thus used in this thesis. In the following we concentrate on baseline results related to the NIST speaker recognition evaluation (SRE) datasets. They are telephony-based and relate arguably to the most appealing use of voice recognition, namely remote recognition over the telephone.

The telephony scenario is one of the most challenging in terms of spoofing and countermeasures since it is entirely unsupervised and is thus particularly prone to spoofing attacks.

[1] The work related to physical access scenario is reported in TABULA RASA deliverables

5.1.1 ASV systems setup

In all experiments the NIST'04 dataset is used for background data, e.g. that used for learning the universal background model (UBM) and that used in the application of score-normalization, nuisance attribute projection (NAP) and factor analysis (FA), while NIST'08 dataset is also used to train the total variability matrix T in the i-vector scheme. The NIST'05 dataset is used for development whereas the NIST'06 dataset is used for evaluation[2].

Features are composed of 16 linear frequency cepstral coefficients (LFCCs), their first derivatives and delta energy, thereby producing a feature vector with 33 coefficients which are computed from Hamming windowed frames of 20 msec and with a frame rate of 10 msec. Voice activity detection is then applied using energy coefficients which are first normalised to fit a zero-mean and unity-variance distribution. They are used to train a three-component GMM which aims to classify acoustic frames into speech/non-speech according to acoustic energy. Speaker modelling is applied only to speech frames; non-speech frames are discarded.

A total of six different systems were assessed and optimised in a similar fashion to the work reported in [72]. All systems were tested with and without the application of T-norm (except for the i-vector system which uses S-norm) using an impostor cohort from the NIST'04 database. All systems have their roots in the standard GMM:

- **GMM-UBM:** The classical system is the GMM-UBM approach with T-norm likelihood score normalization. The UBM model is trained with an EM algorithm. Speaker models are adapted from the UBM via the maximum a posteriori (MAP) adaptation of the GMM mean vectors. Diagonal covariance matrices are not adapted. A top ten component selection is used for likelihood computation. The GMM-UBM system is used as a basis for all other systems.

- **GMM supervector linear kernel (GSL):** The GSL system uses an SVM classifier which is applied to GMM supervectors. The approach is based on that described in [39]. The GSL system is known to outperform other related approaches such as the generalized linear discriminant sequence kernel (GLDS). Supervectors come directly from the GMM-UBM system.

- **GMM supervector linear kernel + nuisance attribute projec-**

[2]Some of this work was conducted through external collaboration with Swansea University

tion (GSL-NAP): The GSL-NAP system is identical to the GSL system but is enhanced with nuisance attribute projection to attenuate intersession (interchannel) variability [40]. Performance is dependent on the rank of the NAP matrix. All experiments reported here correspond to NAP matrices of rank 40 and are learned on the full male and female subsets of the NIST'04 database.

- **Factor analysis (FA):** A FA-based system is also proposed. It is implemented by following the novel latent symmetrical approach [135] to Kenny's original work presented in [101]. This new strategy allows the results (GMM models without session effects) to be used directly in a SVM classifier, among other advantages. All work reported here relates to an intersession matrix corresponding to the 40 most significant eigenchannels.

- **GSL with FA supervectors (GSL-FA):** The GSL-FA system aims to exploit the complementarity in the FA and GSL-NAP systems. The system, reported in [72] is a discriminative SVM approach applied to mean supervectors evaluated in the factor analysis framework.

- **i-vectors with probabilistic linear discriminant analysis compensation (IV-PLDA):** The i-vector system, the current state of the art in speaker verification [59], uses FA to model session and speaker variability at the front-end by means of a so-called total variability matrix. The setup involves mixtures of 1024 Gaussian components and i-vectors with 400 dimensions. The total variability matrix estimation and i-vector extraction is performed using the version 2.0 of the ALIZE toolkit [32] and the LIA-RAL framework [31]. Unwanted variability is handled through Probabilistic Linear Discriminant Analysis (PLDA) compensation [118] with length normalization [81].

For the i-vector scheme, due to the significant amount of data necessary to estimate the total variability matrix T, the NIST'06 dataset was used as background data during development and the NIST'05 dataset was used during evaluation. In both cases the background datasets were augmented with the NIST'04 and NIST'08 datasets. In both cases, matrices are estimated with approximately 11,000 utterances from 900 speakers, while independence between development and evaluation experiments is always respected.

5.1.2 Protocols & metrics

We aim to assess the effect of spoofing on a range of systems based on recent developments in the field of automatic speaker recognition. All of them lead to state-of-the-art performance as judged by the series of NIST SREs and are all based upon the ALIZE toolkit [33]. The inclusion of multiple baseline systems in the case of speaker recognition is motivated by the likely impact of different channel compensation algorithms which may be of assistance to a would-be spoofer.

As is common practice, separate systems are independently optimised for both male and female data subsets. We focus only on the male subset in this thesis. All experiments relate to the core condition (1conv4w-1conv4w) which involves approximately 2.5 minutes of data for model training and testing and all systems are optimised according to the standard EER metric with dynamic performance assessed according to the standard detection error trade-off (DET) plots. Note that, in the case of the NIST SRE datasets, this is in contrast to convention which dictates optimisation according to the minimum decision cost function (minDCF).

5.2 Specifications for vulnerability assessment

This section describes the experimental framework used to evaluate spoofing in this thesis. For the voice biometric we will investigate the following attacks:

1. **Artificial signals:** which can be used on their own or injected into an attackers natural voice signal in order to boost the system score.
2. **Voice conversion:** aims to transform an attacker's voice toward that of a client. We will concentrate mostly on this form of attack.
3. **Voice synthesis:** will only be addressed as a proof of concept and to compare its efficacy with other attacks.

Replay attacks are not assessed in the case of mobile/telephony data since the somewhat artificial nature of the sensor-level attacks to be investigated with NIST SRE datasets means that replay attacks will be impossible to detect; they will not differ from a regular genuine client trial. Thus, replay attacks are not assessed for the mobile/telephony scenario. This section describes the setup of the spoofing attacks as well as the design of the spoofing datasets and the protocols for licit transactions and spoofing assessment.

5.2.1 Spoofing attacks description & setup

Spoofing systems setup are carried out under two important assumptions. While it is admittedly not representative of real scenarios, we assess countermeasure performance in a worst case scenario, where the attacker/spoofer has full prior knowledge of the ASV system i.e. technology used, ASV system configuration, etc. On the other hand, we keep the data used to learn the spoofing system (i.e. training and background data) independent from the data used in the targeted ASV systems.

5.2.1.1 Artificial signals

Artificial signals are generated as illustrated in Figure 4.2. The ASV system used for spoofing is the GMM-UBM without score normalization[3] and with the same configuration presented in Section 5.1.1, but with an UBM trained on NIST SRE'08 data instead of NIST SRE'04.

The speech signal X is divided into frames of 20 msec with a frame overlap of 10 msec. ASV scores are generated for each frame in order to identify the short interval $T = \{c_1, \ldots, c_n\}$ in X with the highest average score. We conducted experiments with values of n between 1 and 20 frames and observed good results with a value of $n = 5$.

The genetic algorithm was implemented using the MATLAB Global Optimization Toolbox V3.3.1. Except for the maximum number of generations which is set to 50, we used MATLAB's default configuration. For detailed information about the system's setup, readers are referred to [12].

5.2.1.2 Voice conversion

All work reported on converted voices was conducted with our implementation of the Gaussian dependent filtering (GD-GMM) approach originally proposed in [133] and described in Section 3.2.4.2.

Due to the assumption of the worst case scenario, the front-end processing used in voice conversion is thus exactly the same as that used for ASV systems. The filtering model g_{fil} and filter $H_x(f)$ uses 19 LPCC and LPC (alpha coefficients) respectively, calculated with SPRO.

[3]Note that this ASV system configuration is used for artificial signal generation, and may or may not differ from the baseline's setup in this thesis

5.2.1.3 Speech synthesis

Spoofing attacks with speech synthesis were implemented using the hidden Markov model (HMM)-based Speech Synthesis System (HTS)[4] and the specific approach described in [204].

Parametrisations include STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) features, Mel-cepstrum coefficients and the logarithm of the fundamental frequency ($\log F_0$) along with their delta and acceleration coefficients. Acoustic spectral characteristics and duration probabilities are modelled using multispace distribution hidden semi-Markov models (MSD-HSMM) [171].

Speaker dependent excitation, spectral and duration models are adapted from corresponding independent models according to a speaker adaptation strategy referred to as constrained structural maximum a posteriori linear regression (CSMAPLR) [203].

Finally, time domain signals are synthesized using a vocoder based on Mel-logarithmic spectrum approximation (MLSA) filters. They correspond to STRAIGHT Mel-cepstral coefficients and are driven by a mixed excitation signal and waveforms reconstructed using the pitch synchronous overlap add (PSOLA) method [140].

5.2.2 Spoofing datasets

The mobile/telephony scenario will be addressed in this thesis using NIST Speaker Recognition Datasets[5] using exactly the same datasets as used for baseline evaluations as reported in Section 5.1. In summary the NIST'04 and NIST'08 datasets are used for background, normalisation or session modelling, the NIST'05 datasets are used for development and the NIST'06 dataset is used for evaluation. Further details regarding the specification of each dataset is freely available from NIST's website. All data used to effect spoofing attacks will come from the same datasets.

The only difference between the data used for baseline experiments, reported in Section 5.1, and the data to be used for spoofing relates to the use of a different experimental condition. Instead of the 1conv4w condition used for baseline experiments, spoofing and related experiments will be performed on

[4] http://hts.sp.nitech.ac.jp/
[5] http://www.itl.nist.gov/iad/mig/tests/spk/

NIST SRE datasets - 8conv4w-1conv4w - male subset		
	Development (NIST'05)	Evaluation (NIST'06)
Baseline	201/984/8962	298/1344/12648
Artificial Signals	201/984/201	298/1344/298
Voice Conversion	As for baseline	
Speech Synthesis	As for baseline	

Table 5.1: Size of the datasets used for spoofing assessment in the case of the mobile/telephony scenario (NIST datasets). Figures illustrate the number of clients / genuine client trials / and impostor trials.

the 8conv4w condition which provides multiple sessions for each speaker. It contains voice recordings from 201 and 298 speakers for NIST'05 and NIST'06 male subsets, respectively. There are 8 sessions for each speaker giving a total of $499 \times 8 = 3992$ recordings of approximately 2.5 minutes in duration. Those sessions not used for testing are used to effect spoofing attacks for all impostor trials in the standard NIST protocols.

In addition to the new baseline dataset (without spoofing), three new datasets will be generated where all impostor test segments are replaced with spoofed versions coming from artificial signals, voice conversion or voice synthesis.

As mentioned in Section 4.3.2, under the assumption that all system elements can be modelled as different, linear processes (as is the case for the three attacks investigated), then spoofing, acquisition and transmission (channel) is equivalent to acquisition and transmission followed by spoofing. Under the assumption of linearity, spoofing work involving NIST SRE datasets can therefore be performed directly on the exact same, pre-recorded datasets, without re-recording, while still being consistent with sensor-level spoofing.

5.2.3 Protocol for licit biometric transactions

Except for the use of the 8conv4w condition instead of the 1conv4w condition, the protocols for licit biometric transactions are exactly the same as those for all baseline results reported in Section 5.1. Once again, the NIST'04 and NIST'08 datasets are used for background data, the NIST'05 dataset is used for development and the NIST'06 dataset is used for evaluation. All experiments relate to the core condition which involves approximately 2.5 minutes of training and testing data.

The new protocols/conditions result in a slightly reduced number of speakers

and recordings than used for previously reported baseline experiments. A summary of the dataset sizes illustrating the number of clients, genuine client trials and impostor trials is illustrated in Table 5.1 for both development and evaluation datasets. The development set contains 201 male clients whereas the evaluation set contains 298 male clients. These numbers are consistent for the baseline protocol and the three spoofing protocols (described below). The precise protocol is non-exhaustive and exactly as defined by NIST.

5.2.4 Protocol for spoofing attacks

The protocols for spoofing assessment are identical to the licit protocol only that, for each impostor access attempt, the test sample is treated or replaced according to the given spoofing attack. The protocols for the NIST SRE datasets differ slightly and are described below.

For the NIST SRE datasets, any data used to effect spoofing comes from one of the sessions not used for testing. Of the 8 available, the first is used for testing, thus sessions 2 to 8 can be used for spoofing purposes. Furthermore, any other suitable data, e.g. the NIST 2008 dataset, will be used as independent background data (i.e. for learning a universal background model used in voice conversion). Except for modifications to impostor test segments through spoofing, protocols are exactly the same as described above. Accordingly there is no overlap between the data used to train client models and that used for spoofing.

The number of genuine client trials remains the same as described above, however, the number of impostor trials is dependent on the form of spoofing attack. As illustrated in Table 5.1 the number of impostor trials for artificial signals is greatly reduced (201 c.f. 8962 for development and 298 c.f. 12648 for evaluation). This is because, when comparing a voice model for person A with an impostor test segment B, the impostor test segment is replaced with an artificial signal which is targeted only towards model A and is independent to impostor B.

With voice conversion, however, a transformation is learned which maps the impostor test segment B toward model A while for the case for voice synthesis we use the transcripts of impostor B to obtain a speech sample per impostor trial. It is specific to the test segment and thus the number of impostor tests in this case is the same as that in the baseline. As a result, with artificial signals and speech synthesis spoofing attacks the number of impostor trials is reduced to 1231 male and 1337 female tests for the development set and 1543

System	Development		Evaluation	
	no-norm	norm	no-norm	norm
GMM-UBM	8.2	8.1	9.1	8.6
SGL	7.8	7.8	7.9	8.1
SGL-NAP	5.9	5.9	6.3	6.3
SGL-FA	5.1	5.1	6.1	5.7
FA	4.7	5.1	5.6	5.6
IV-PLDA	4.2	4.3	3.4	3.2

Table 5.2: Equal error rate (EER) scores (%) for each of the six speaker verification systems and for male subset. Results are illustrated for the development (NIST'05) and evaluation/test (NIST'06) datasets.

male and 1843 female tests for the evaluation set.

5.3 Results

Since there are numerous, different approaches to speaker recognition in the literature, a range of systems is utilized: a standard Gaussian mixture model with universal background model (GMM-UBM), a GMM supervector linear kernel (GSL) system, a GMM supervector linear kernel system with nuisance attribute projection (GSL-NAP), a factor analysis (FA) system, a GSL system with FA supervectors (GSL-FA) and an i-vector system with probabilistic linear discriminant analysis compensation (IV-PLDA). We also considered three different spoofing attacks, including artificial signals, voice conversion and synthesized speech defined in Section 5.2.1.

5.3.1 Baseline

We ran a series of experiments designed to compare the performance of the GMM-UBM, GSL, GSL-NAP, GSL-FA, FA and IV-PLDA systems for male subsets. All systems were optimised independently on the development set and were then applied without modification to the evaluation set following the protocols in Section 5.1. The EERs for each system are presented in Table 5.2 and show the evolution in performance with different approaches to compensate for intersession variation.

For the male subset the best performing IV-PLDA system (judged from the

development set) gives an EER of 3.2% on the evaluation set. This compares well to the GMM-UBM system where the respective EER is 8.6%. Upon comparison of these results to those reported in the most recent NIST SRE campaigns we note that the tested systems represent the state-of-the-art in current automatic speaker recognition technology and are therefore suitable candidates for assessing the threat from spoofing and for testing countermeasure developments.

DET plots for the evaluation set are illustrated in Figure 5.1. They relate to the ASV systems for the male subset without and with score normalisation, illustrated in Figures 5.1(a) and 5.1(b), respectively. From the analysis of the profiles in Figure 5.1 and the results in Table 5.2 we observe that the differences in performance between systems with an without score normalisation are not significant.

It is worth noting that the optimised ASV configurations remain the same for experiments in the following sections and are not necessarily the system configurations which give the best performance on the licit transaction protocols.

5.3.2 Spoofing

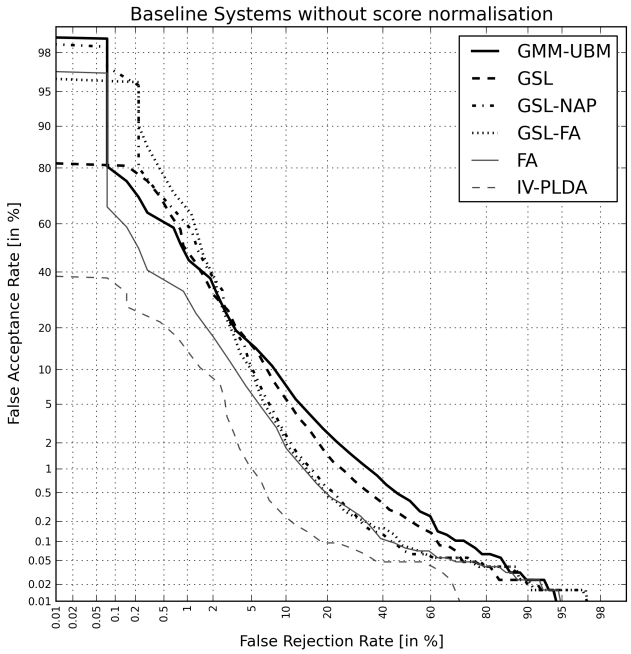
Results are illustrated through classical detection error trade-off (DET) profiles, through a summary of false acceptance rates (FARs) for fixed false rejection rate (FRR) and through illustrations of client, impostor and spoofing score distributions.

Experiments include the high-effort attacks defined in Section 5.2.1 and an example of low-effort attack with white noise proposed in Section 4.2.2.

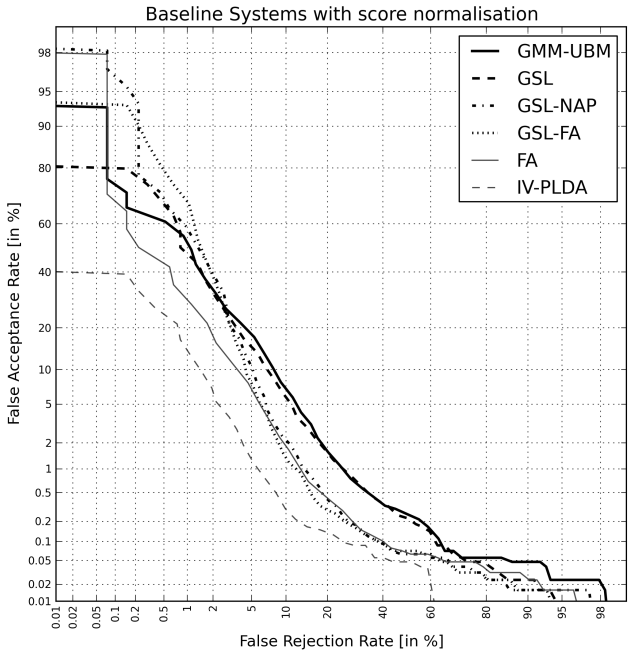
While in this section we present results for the six different systems we concentrate on the GMM-UBM and IV-PLDA systems. The former is arguably the most popular approach to speaker recognition whereas the latter is representative of the state-of-the-art and gives the best performance among the six systems tested, according to results in Section 5.3.1.

5.3.2.1 High-level effort attacks

Results for ASV systems with and without score normalisation are illustrated by mean of a collection of FARs presented in Table 5.3(b) and Table 5.3(a), respectively. The tables report results on ASV systems performance for licit transactions and under spoofing attacks.



(a) Baseline systems without score normalisation



(b) Baseline systems with score normalisation

Figure 5.1: Detection error trade-off (DET) profiles for male subsets of the evaluation/test NIST'06 dataset. ASV systems are evaluated with and without score normalisation

FARs are calculated by fixing the FRR of each ASV system to its baseline EER value (%). Results of the performance of the six ASV systems for licit transactions, illustrated in the first column of Tables 5.3, are thus directly comparable with the EER values presented in Tables 5.2. We report similar performances for each of the 6 analysed ASV systems, despite of the training protocols coming from 1conv4w (Section 5.1) or 8conv4w (Section 5.2). The IV-PLDA system gives again the best performance with an FAR of 3%.

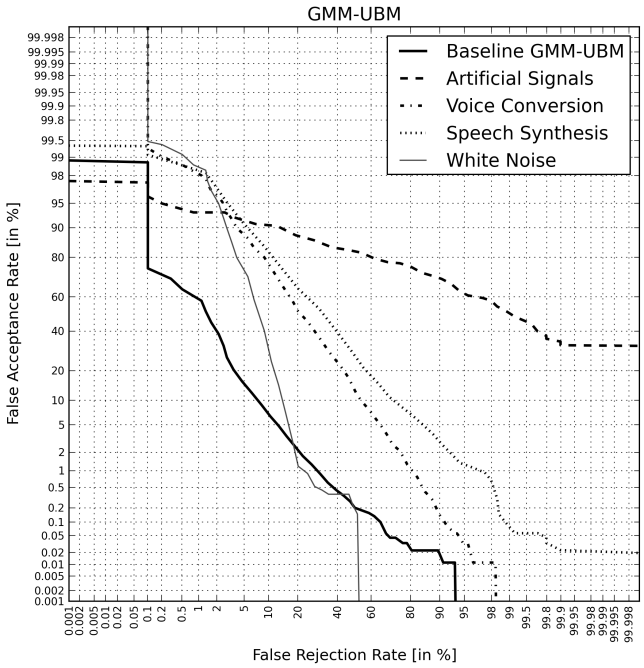
From the considered attacks, the ones with converted voices appear as the most serious threat for most of the ASV systems; baseline FARs between 3% and 9% increase to values between 71% and 94% for all the ASV systems analysed. Smaller, but still significant increases in the FAR are reported for voice synthesis, with increases in false acceptances between 36% to 87%.

For the latter, significant degradations are observed in all cases, except for the artificial signal attacks and the three GSL-based and IV-PLDA systems. This is not a surprise since the the GSL supervectors model speech at the GMM component level, whereas the artificial signal attacks, generated by using a standard GMM-UBM system, target ASV systems at the feature level. Moreover, GMM-UBM and FA systems are relatively more robust than GMM supervector-based approaches to attacks with voice conversion (being FA the most roust of the six systems) , while the opposite seems to happen with spoofing attacks with speech synthesis.

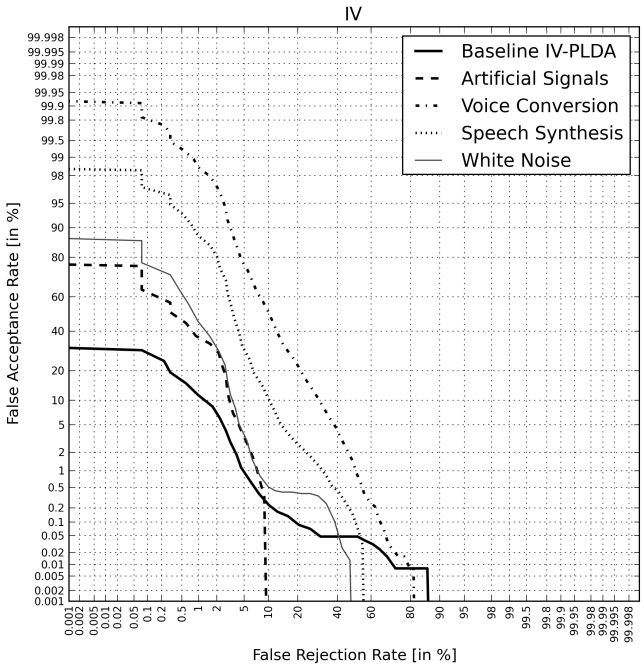
The ambiguous impact of score normalisation is also visible in Table 5.3(b). For IV-PLDA ASV systems, symmetric score normalisation (S-norm) mostly helped to decrease FAR values in the face of spoofing, e.g. for attacks with speech synthesis the FAR decreased from 54% to 36%. In contrast, in some cases, e.g. for artificial signals or speech synthesis and GSL systems, the FAR increased after applying score normalisation. The influence of score normalisation in the face of spoofing is still unclear and subject for further research.

DET plots for the mobile/telephony scenario are presented in Figure 5.2. Plots are illustrated for GMM-UBM system (a) and IV-PLDA system (b), both without score normalisation.

Compared to the baseline, voice conversion and speech synthesis both provoke increases in the false acceptance rate (FAR) across the full range of thresholds and the threat from voice synthesis is marginally greater than that from voice conversion for GMM-UBM systems, and vice-versa for IV-PLDA systems. Artificial signals also provoke increases in the FAR for the GMM-UBM system (for which it has being optimised). Below false rejection rates (FRRs) of 3%, artificial signals give the greatest increase in FAR. At lower values of FRR the



(a) GMM-UBM



(b) IV-PLDA

Figure 5.2: Speaker verification performance using GMM-UBM and IV-PLDA systems for the mobile/telephony scenario. Profiles shown for the baseline and different spoofing attacks.

(a) ASV systems without score normalisation.

System/Attack	-	AS	VC	SS	WN
GMM-UBM	9.1	93	78	87	53
SGL	7.9	2	92	41	2
SGL-NAP	6.3	8	88	55	2
SGL-FA	6.1	1	90	39	4
FA	5.6	73	71	77	36
IV-PLDA	3.0	11	94	54	13

(b) ASV systems with score normalisation.

System/Attack	-	AS	VC	SS	WN
GMM-UBM	8.6	70	91	72	4
SGL	8.1	2	92	42	3
SGL-NAP	6.3	21	88	57	4
SGL-FA	5.7	19	73	56	5
FA	5.6	38	83	59	8
IV-PLDA	2.9	16	85	36	1

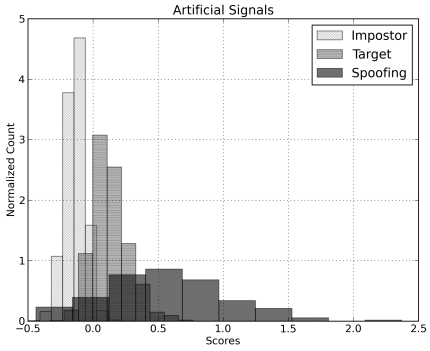
Table 5.3: False acceptance rate (FAR) scores (%) for fixed false rejection rates (FRR) set to the EER baseline for each of the six speaker verification systems in the mobile/telephony scenario and for artificial signals (AS), voice conversion (VC), speech synthesis (SS) and white noise (WN) spoofing attacks. The first column correspond to ASV performance with no attacks (licit transactions).

increase in FAR from artificial signals is smaller than that for voice conversion and voice synthesis.

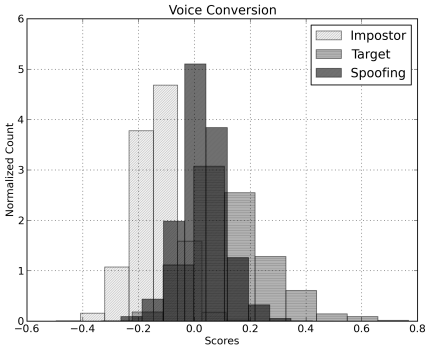
From Figure 5.2(b) we note that this particular configuration of artificial signals generation does not represent a threat to IV-PLDA systems, however, as mentioned in Section 5.2.1.1, initial experiment suggest that artificial signals trained with a similar (IV-PLDA) system represent a serious threat.

Client, impostor and spoofing distributions for each of the three attacks are illustrated in Figures 5.3 and Figures 5.4. They correspond to the DET profiles illustrated in Figures 5.2(a) and 5.2(b) for the GMM-UBM and IV-PLDA systems, respectively.

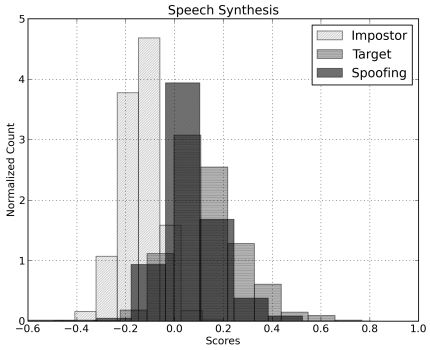
All plots show that the distribution of spoofed scores overlaps more with the distribution of target, client scores than the impostor distribution. For the GMM-UBM system, the distribution of spoofing scores for artificial signals



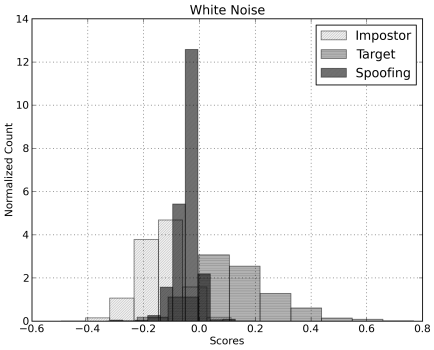
(a) Artificial signals



(b) Voice conversion



(c) Voice synthesis



(d) White noise

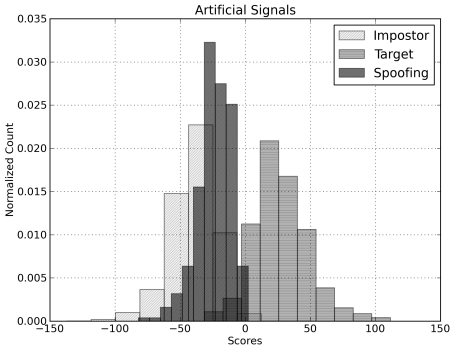
Figure 5.3: Client, impostor and spoofing score distributions for the GMM ASV systems and each of the three high-effort spoofing attacks and attack with white noise for the mobile/telephony scenario.

has a higher variance and some especially high scores, while for the IV-PLDA case such distribution does not overlap significantly the genuine distribution score.

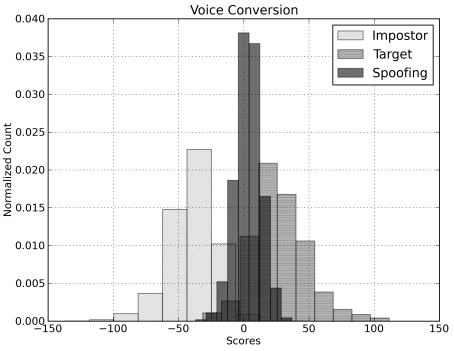
These plots confirms that spoofing can bias scores toward the target distribution meaning it is then more difficult to differentiate between genuine clients and spoofed signals.

5.3.2.2 An example of low-effort attack

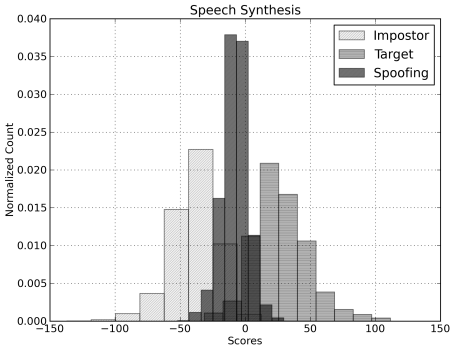
An example of low-effort attack is introduced in this thesis to motivate further research on the hypothetical threat of low-effort spoofing attacks (Section 4.2.2)



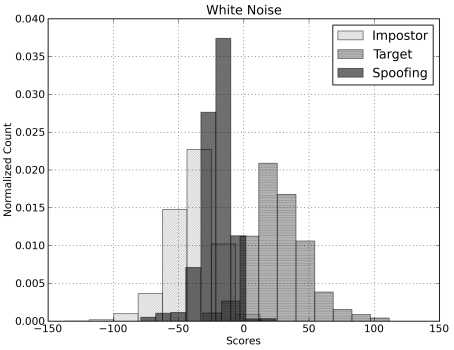
(a) Artificial signals



(b) Voice conversion



(c) Voice synthesis



(d) White noise

Figure 5.4: Client, impostor and spoofing score distributions for the IV-PLDA ASV systems and each of the four spoofing attacks for the mobile/telephony scenario.

We report a small experiment in which the impostor trial utterances are replaced by signals containing only white noise and therefore tested against the six ASV systems. Results related with this experiment are showed in the fourth column in Tables 5.3 for the six ASV systems and Figures 5.2(a) and 5.3(d) and Figures 5.2(b) and 5.4(d) for GMM-UBM and IV-PLDA systems, respectively.

Table 5.3(a) shows that non SGL-based approaches are indeed vulnerable to attacks with white noise, while on the use of score normalisation in general acts as a good countermeasure against such a threat. Below FRRs of 10% and 2% for GMM-UBM (Figure 5.2(a)) and IV-PLDA (Figure 5.2(b)), white noise attacks give increases in FAR above 30%.

5.3.3 Discussion

As mentioned in Section 4.2.3, the approach to artificial signals proposed in this thesis is a system-dependent attack i.e. the effectiveness of the attack depends on the similarities between the targeted ASV system and the ASV system used to synthesize the attack. However, the results reported in Table 5.3 show that artificial signals provokes significant increases in the FAR of a variety of ASV systems, specially for those with normalised scores, even thought they were generated only by using a GMM-UBM system without score normalisation.

Furthermore, our work reported in [12] shows that artificial signals are also effective when the configurations of the targeted and spoofing ASV systems mismatch[6]. Future work should evaluate the risk of attacks with artificial signals that are synthesized by using more sophisticated ASV systems i.e. i-vectors + PLDA post-processing.

Similar to previous work [106], the experimental results in Table 5.3 also suggest that ASV systems employing intersession compensation might be intrinsically more robust to voice conversion attacks. These results are in contrast to our first hypothesis that channel compensation approaches may be of assistance to a would-be spoofer by mitigating the 'channel shift' or the artifact produced by speech synthesis or voice conversion systems. This behaviour is still unexplained and subject to future research.

We note that ASV systems with normalised scores are effective in detecting attacks with white noise. In fact, score normalization combines two powerful heuristics against spoofing attacks. First, we observe from Equation 2.1 that if X results from a spoofing attack designed to produce high scores (or values of $L_\lambda(X)$) independently of the speaker (or model λ), it will produce also a high value of μ_λ and consequently a low $\widetilde{L}_\lambda(X)$; thus, this heuristic acts as a good countermeasure against generalised attacks (Section 4.2.3). Second, the term σ_λ compensates against random or noisy-like signals which produces a wide range of scores and which could be used to perform brute-force attacks.

Nevertheless, generalised attacks still as the potential to bypass score normalization. For instance, if X is the result of a spoofing attack such that it produces the same or very similar scores i.e. scores in a narrow range independently of the speaker, then $(L_\lambda(X) - \mu_\lambda) \approx 0$ and $\sigma_\lambda \approx 0$. If we assume that the distribution of scores is symmetric, then will be similar that tossing

[6] we remind the reader that in this thesis the targeted and spoofing ASV systems share the same configuration, but use different training data

a coin.

Indeed, to generate a signal of this nature requires prior knowledge of the targeted recognition system. Although the design of such a signal it is out of the scope of this thesis, future work should include the analysis of more complex signals which could bypass score normalization.

Finally, even thought there is no previous work on generalised spoofing attacks, inferences can be made based on observations from experimental work from different sources. Results in Section 5.3.2.2 suggest that an audio signals containing purely white noise can be marginally a better attack than naive impostor, even though the increase in false acceptances is not significant, we think that these results suffice to encourage further research.

Chapter 6

Countermeasures: Fundamentals

The problem of reliable biometric identity verification involves biometric systems, spoofing, countermeasures and the relation between them. Due mainly to the novelty of this problem, fundamental questions are still unclear. For instance, is there a need to reformulate the problem of biometric identity verification considering the new paradigm of spoofing and countermeasures? If that is the case, how to proceed? Is this still a binary client/impostor classification problem?

From the point of view of the author of this thesis, these fundamental issues has been disregarded or at least not properly addressed by the research community. For instance, as trivial as the latter question may appear, it still remains without consensus among researches from different areas. Part of the contribution of this thesis is to provide an overview including current approaches and specially to extend it by presenting our own, novel approach to address this problem.

Another issue related specifically with countermeasure development and pointed out in Section 3.4 refers to the weaknesses in countermeasure evaluations. In general, the research on anti-spoofing tends to focus on the spoofing detection itself and omit to make the link with the recognition system, however, in practice a countermeasure will never be used stand-alone. This thesis thus addresses the problem of combining biometric systems and countermeasures, which is an unavoidable stage in future countermeasure evaluations.

The contributions of this chapter are related with the the issues commented above. Section 6.1 presents different approaches to the formulation of this new problem and highlight the approach followed in this thesis while Section 6.2 presents our own analysis of countermeasure integration.

6.1 Approaches to problem formulation

Coming back to the latter question, what would be the basis to define the number and type of classes for this problem? Are they unique, predefined and intrinsic to the problem or are they ruled by application requirements?

Traditionally biometric identity verification is defined as a two-class problem, but, from the application point of view, we might want to know if a rejected attempt was the result of a naive impostor or due to an intentional spoofing attempt and therefore three classes should be considered.

In this thesis, we consider the formulation of the problem as a part of the system design problem. Then, different approaches will be more or less suitable depending on different factors, including the state of the art in countermeasures or the potential implementation cost effort in creating new evaluation standards, among others.

Although the remaining explanations in this section are supported mostly by examples for voice, they can be extended also other biometrics.

6.1.1 *Two-class* approach

The first approach to the formulation of the spoofing and countermeasures problem is motivated from the nature of the problem itself, which is a **two-class problem**: person should either be granted the access, or not. Thus, impostor remains impostor independent of whether it is a replay sample, synthesis sample or a random "naive" impostor.

Formulated this way, spoofed impostor samples are considered as intra-class variation of the class *impostor*. Researches in the speech community that support this idea claim that speaker recognizers should be able to handle spoofing in the same way that they handle different kinds of channels and environments using the same system. Speaker verification systems are always evaluated with a wide range of different channels, microphones and noises and in this context spoofed imposture is seen as a similar effect to varying channel or environment.

To this end, our experiments reported in Section 5.3 together with some observations [106] suggest that advanced algorithms such as joint factor analysis [101] may offer some inherent protection from spoofing. One of the outcomes of this approach could be to continue with the traditional pursuit of improved fundamental performance, with extended databases containing spoof-

ing samples and new protocols (metrics for performance evaluation could be reused here) and the possibility to treat spoofing with a powerful background theory such as statistical hypothesis testing.

One of the main reasons why we cannot find examples of this approach in the literature is related to the difficulty to address this problem "all at once". But undoubtedly, the main drawback of this approach is related to the practical aspect.

While different speakers, channels and environments are assumed to be well characterized in current speaker recognition evaluations (i.e. NIST databases), the task of compiling a set of samples that provides a comprehensive characterization of the 'spoofing' concept (or everything that is not the client concept), as is assumed in conventional two-class problems, is believed to be extremely difficult or impossible.

While attacks from impersonation, replay, speech synthesis and voice conversion are all known, there is a high degree of variation in specific algorithms and there are certainly other forms of attack yet to be identified, which makes infeasible the development of the conditional density function for the alternative hypothesis.

This problem does not affect only voice biometrics. For instance, *de Freitas Pereira et al.* [51] showed that state-of-the-art spoofing countermeasures for face recognition do not generalise well to forms of spoofing not considered during development.

While the true extent of spoofing in the context of ASV is yet to be fully understood, and in any case, there still remain the question of how costly would be to properly estimate the conditional density function of the spoofing concept. This question, as other several fundamental ones, is open to research, but taking into account that spoofing comprehends attacks going from impersonation to non-speech, artificial audio signals, this task is, at least, challenging.

Due to the state of the art in countermeasures, it is likely to expect a significant percentage of previously unseen attacks. Therefore, the problem of unknown attacks is a major problem in the spoofing context and should be treated accordingly.

6.1.2 *Three-class* approach

A second approach is the formulation of a **three-class or ternary classification problem** [47], consisting in clients, impostor and spoofing attacks.

One of the advantages of the three-class thinking is that gives more detailed diagnostic information, which as mentioned before could be an application requirement, but most importantly is the fact that this approach allows the design of spoofing countermeasures independently of the biometric system. Thus, independent countermeasures have the advantage of being incorporated easily into any existing biometric system and to specifically detect spoofing attempts.

A straightforward way to address the problem of spoofing and countermeasures formulated as a three-class problem is to combine a biometric system with a "generalised" spoofing attack detector, in which the detector (a binary classifier) is fed with features from legitimate samples as positives and attacks from different sources as negatives. Some work [90], not necessarily related to spoofing, suggest that this approach is appropriate as long as the negative samples are representative of the concept i.e. the universe of possible attacks.

Another advantage of the three-class view is that current methods, protocols and metrics for evaluation of spoofing and countermeasures [48], including the one used in this thesis (Section 3.3.1) are designed based in a formulation of a three-class problem.

However, this formulation still does not handle unseen attacks, is more likely that the spoofing samples used to train the classifier may not represent the negative concept uniformly and may involve human bias. Together with the two-class approach, they ignore the reductionist outlook adopted for the research community to face the problem of countermeasure development.

Basically most of the work in the literature is related to the development of countermeasures designed to detect a specific attack following a relatively simple methodology: first, to analyse/investigate/study the specific attack in order to find and to extract characteristic intrinsic (that describes) to the attacks and which also differs from the real accesses and then to model them, usually by using a discriminative approach (i.e. two-class SVM classifier, with negatives generated by the researcher).

6.1.3 *Multi-class* approach

The easiest way to propose a framework which includes also the work done so far is by **considering several classes**, but with two differences from conventional multiclass problems.

First, in the problem of reliable recognition only one class is of interest i.e. given a biometric sample, we want to know if it belong to the class *client* or not; and second is the fact that in practice it is likely to expect observations which do not belong to any of the classes defined in the system.

That said, we present what in our understanding are the ways to handle spoofing, which are described in the rest of this section.

6.1.3.1 Conventional *multi-class* approach

The first approach is still to consider a **conventional multi-class classification problem**. How to design, under this context, a generalised countermeasure when in the problem formulation all classes (attacks) are assumed to be known?.

One solution is to develop a modular system. Adding new detection rules or coupling new classifiers to a system in response to new attacks is a common practice in spam filtering and in intrusion detection in computer systems, this is what we call the "anti-virus-style framework" and is also the approach observed in recent efforts to develop generalised countermeasures to deal with the variation in possible attacks [46].

The idea is: every time a new specific attack or spoofing indicator is known[1], a new classifier is optimised to detect that threat and therefore coupled to the ensemble of classifiers. The same technique may also be used to enhance robustness to a single, specific attack e.g. the combination of motion and texture analysis for face anti-spoofing [110].

Binary spoofing detectors are generally, independent of the biometric system and typically trained using both genuine data (negative samples) and examples of spoofed data (positive samples).

The main drawback of such an approach is the over fitting to spoofing attacks seen in the training data and thus the lack of generalization to attacks

[1]For the conventional multi-class problems it is assumed that the new attacks are detected off-line.

previously unseen [51]. While additional classifiers can be trained and integrated when new attacks are identified, clearly this leads to increased system complexity.

6.1.3.2 *One-class* approach

The second approach is to formulate a **one-class classification (OCC) problem**. The main difference between OCC and two-class and multiclass problems is that the former groups all the techniques and methods that model objects from a single class, which makes it specially suitable for cases where only one class of data is available, and others are too expensive to acquire or too difficult to characterize. For this problem, the one-class formulation in its pure form is mainly theoretical and unlikely to be implemented in practice, since it disregards the use of most of prior knowledge as well as the exiting biometric systems and countermeasures.

6.1.3.3 *Multi-class* approach with outliers detection

A third approach first proposed in this thesis is to formulate the problem of reliable biometric verification as a **multiclass problem with outliers detection** [163]. The idea is to provide to the conventional multiclass problem the option to handle test samples belonging to an new, unknown class e.g. by adding one (or more) one-class classifier/s to the system.

This approach, also applied for other fields such as intrusion detection, fraud detection fault detection in manufacturing[2], is also denoted as a multiclass formulation with reject option, which consider the outliers belonging to *reject* class [25, 185]. Since for the author of this thesis to have another, *reject* class is opposite to the notion of one-class classification, we keep in this thesis the first notation.

This latter approach is a trade-off between the first two approaches and it is conservative with respect to previous ones in the sense that makes use of as much previous knowledge as we can have and is modular with respect to existing biometrics systems and countermeasures and also addresses the problem of unknown attacks, which makes it the approach adopted in this thesis.

[2]Outliers detection is applied in whatever application where a new class is expected, here we highlighted the applications that resemble spoofing

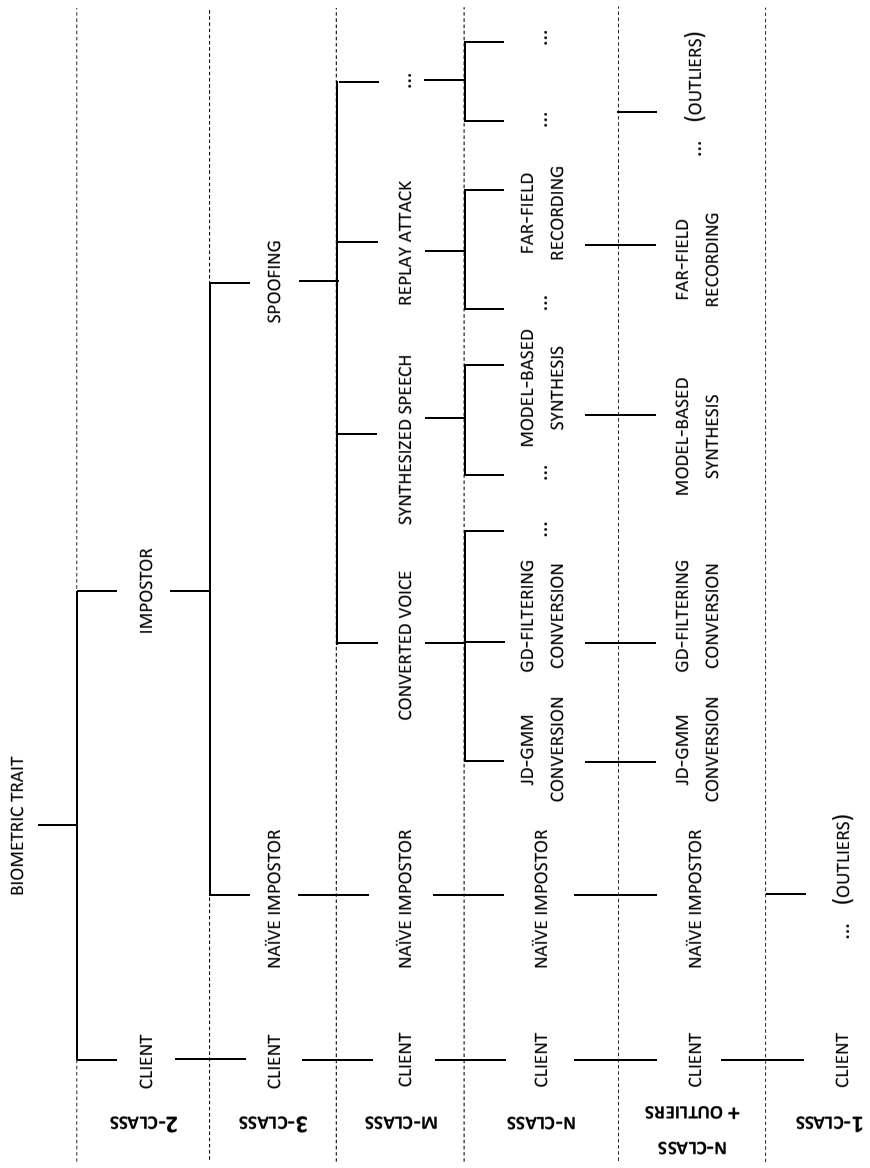


Figure 6.1: Different approaches to formulate the problem of reliable biometric verification, initially formulated as a client/impostor binary classification problem. The class *impostor* is therefore increasingly split with the growing number of classes considered in the formulation of the problem.

Surprisingly, although the managements of outliers seems a natural step to consider, as far as we know this is the first and only work who propose the detection of outliers in the context of biometric spoofing.

Figure 6.1 illustrates the different approaches described in this section. In this thesis we evaluate the problem of spoofing and countermeasures combining a maximum of two systems, then the problem is formulated either as a 3-class problem (ASV system + spoofing countermeasure) or either as a 2-class problem with outliers detection (ASV system + generalised spoofing countermeasure). Combination of more than two systems is addressed in Section 6.2 but not evaluated in this thesis.

6.2 Combining biometric systems and countermeasures

As stated in Section 6.1, most approaches to a solution involve the consideration of several classes which are usually implemented by a ensemble of systems (classifiers), being the simplest approach the use of a biometric system coupled with a countermeasure.

Combination of systems is unavoidable. A countermeasure is basically an auxiliary system designed to increase the robustness of biometric systems against spoofing attacks; thus, a joint operation with a recognition system is the setup where the countermeasure gets its meaning. Nevertheless, as noted in [47], independently of the biometry the research on anti-spoofing tends to focus on the spoofing detection itself and omit to make the link with the recognition system.

Although evaluating a countermeasure as a stand-alone system is valid for comparing the effectiveness of different approaches against spoofing attacks, meaningful results are those obtained from the analysis of both recognition system/s and countermeasure/s working together.

Due of the state of the art in countermeasures development, a recognition system operating in a real-world scenario is expected to count with more than one countermeasure. This is the main motivation for which this thesis stresses the importance of the research of different strategies to combine biometric systems and countermeasures. Such methods, known as **Multiple Classifier Systems** (**MCS**), are considered as one of the most promising research directions in current field of machine learning and pattern recognition [92].

This section aims to summarize the main work related to the subject of this thesis. We acknowledge that this thesis do not report research on MCS. The scope of this section is limited to highlight some aspects related to MCS in the context of spoofing, as well as to provide some insight about the MCS design and specially to point out some issues in current evaluation with respect to this regard.

6.2.1 Previous work on MCS

MCS is a vast and growing research area. Also denoted combination of multiple classifiers, ensemble learning, classifier fusion, mixture of experts, consensus aggregation, voting pool of classifiers, composite classifier systems, classifier ensembles, modular systems, collective recognition, stacked generalization, etc. it is widely accepted that combining multiple classifiers can take advantage of the strength of individual classifiers, avoid their weaknesses and improve classification correctness [169, 47]. There exist a multitude of work related to MCS. The research progress in this topic is well summarized by the successful series of workshops on Multiple Classifiers Systems, conducted yearly from 2000 [108] to 2013 [214].

On the other hand, using MCS in one-class classification is an approach that still awaits proper attention. Most of the work in this topic is application-oriented, e.g. image retrieval [197], monitoring the information network [137] or medicine and biology [205] and, expect for some notable exceptions e.g. [111], there is also a lack of works devoted to the theoretical advances in combination of one-class classifiers. Even though there exist some work that uses one-class, two-class and multiclass approaches to design the decision combination function of an MCS i.e. [87], to the best of our knowledge there is not relevant work related to the combination of one-class, two-class and multiclass classifiers.

MCSs are currently being used in several classification tasks, like multimodal biometric systems, intrusion detection in computer systems, and spam filtering [84, 22]. Recently, when tested against adversarial conditions [23] it has been shown that the robustness of MCS is highly dependent of the combination decision function. In particular, for multimodal systems to overcome only one modality can be enough to fool the system, as discussed in [166, 98, 4]. This has motivated the development of fusion schemes specifically designed to increase the robustness to spoofing i.e. [166, 165] although none of this work employs an algorithm specialized to detect spoofing attacks.

Examples of combination of biometric recognition systems and countermea-

sures are very weird to find. In fact, to the best of our knowledge there are only three publications that address this issue, two related to the work of Marasco et al [124, 125] for fingerprint verification and one example for face modality [47].

In [124], the authors analyses four different score fusion methods for a liveness detector incorporated with a fingerprint matcher, while in [125] the same authors shows a robustness a robustness increase of a multimodal system consisting in three fingerprint and one face modality by incorporating a fingerprint liveness detection algorithm in the combination scheme. The work in [47] analyses three score fusion and one decision rule for a face verification system combined with three different countermeasures, one at a time. Moreover, it present some basis for the evaluation of the final system.

Among the limitations of these examples, common to all of them is the fact that they do not fuse more than one countermeasure at a time, and any of them consider the management of outliers.

6.2.2 Design of MCS in the context of spoofing

This section presents some aspects in the design of MCS in the context of spoofing. Without losing generalization and to simplify the discussion, we assume the scenario most likely to occur in the MCS design process.

Usually, for a given problem we may have a pool of several classifiers at our disposal, consisting in recognition systems and countermeasures. We consider both, systems with and without their thresholds already set. We may note that the number of classifiers is likely not to be the same as the number of classes in the problem. Although in general the "One per Class" (OPC) approach is the most common choice, in our definition of the problem the number of classes is assumed undefined.

Besides, we may want to add to the MCS more than one recognition system, more than one countermeasure to detect a particular attack, countermeasures that detect a group of attacks or multiple one-class classifiers. The resulting MCS will differ from the multimodal systems in the sense that the former combine experts which work with the same biometric modality and accordingly all the classifiers consist in one, common input i.e. the biometric sample. Finally, since we consider critical the management of outliers, we must consider at least one one-class classifier in the MCS design.

According to [167], a MCS can be characterized by its architecture/topology,

the number and type of the base classifiers (the classifier ensemble) and the decision combination function (the fuser).

Schemes of MCS can be grouped according to their topology into three main categories [92] named serial, parallel, and hierarchical/hybrid. Figure 6.2 present examples from the different topologies. For parallel topology the classifiers are invoked individually and merged by a single combination function. In serial topology the classifiers are invoked in sequence, with each classifier producing a reduced set of possible classes. In hierarchical topology, the different classifiers are combined into a tree-like structure.

One key component of the design of a MCS is related to the decision function or fusion strategy. Although fusion can be carried out also at the data level and feature level [92], researches usually relate MCS design to the combination of classifiers at the output level, which is the most studied and most rewarding among the three approaches.

The output of a classifier can take many forms. A formal classification, introduced by [202], classifies them into abstract form, rank level and measurement (or confidence) level. At abstract form, the classifier only outputs an unique class. At rank level, the classifier outputs a ranked list of classes, with the class ranked first being the first choice. At measurement level, the classifier assigns a value depicting the belief or probability that the classifier has of the input value belonging to each class. In a verification framework like the one in our case, it is not likely to find examples of rank-level output.

Also, it is worth noting that if we assume the ensemble consist only with one-class and two-class classifiers with abstract from output (thresholds already set), then the cascade topology is equivalent to a parallel one with a logic AND fusion rule. Still, for this topology there is an option to jointly optimized thresholds all systems [121].

The classifier ensemble refers to the type and number of base classifiers and is related with the notion of *diversity* [112]. To this end, we note that unlike fusion of two biometric recognition experts, our task at hand requires fusing of two discordant systems. A verification and an anti-spoofing system are of different nature and have antagonistic criteria for taking a decision. Related analysis of the ensemble to take into account in future research includes the consideration modular aggregation [73], the fact that in general one class classifiers are less accurate than their two-class homologous, among others.

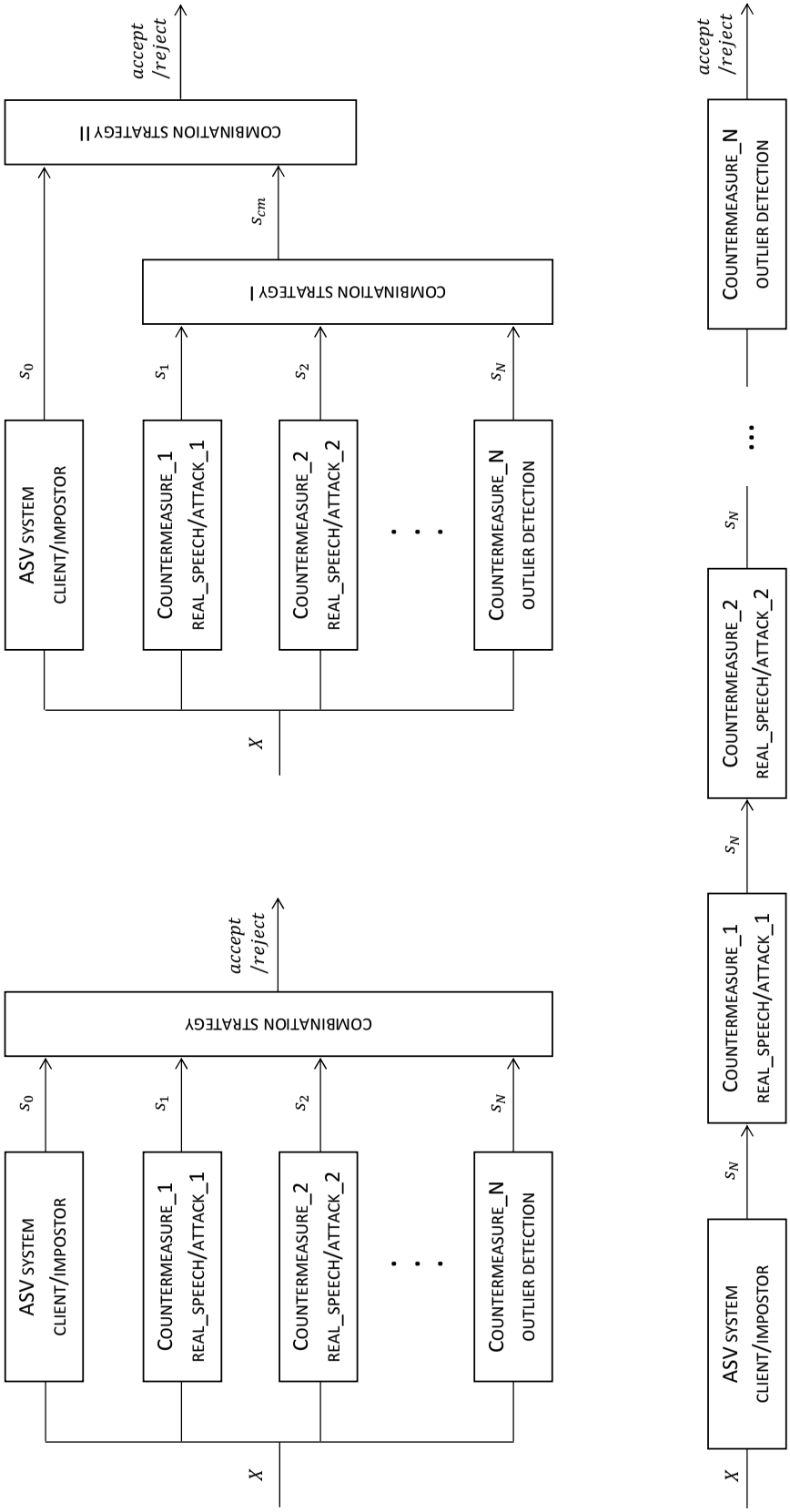
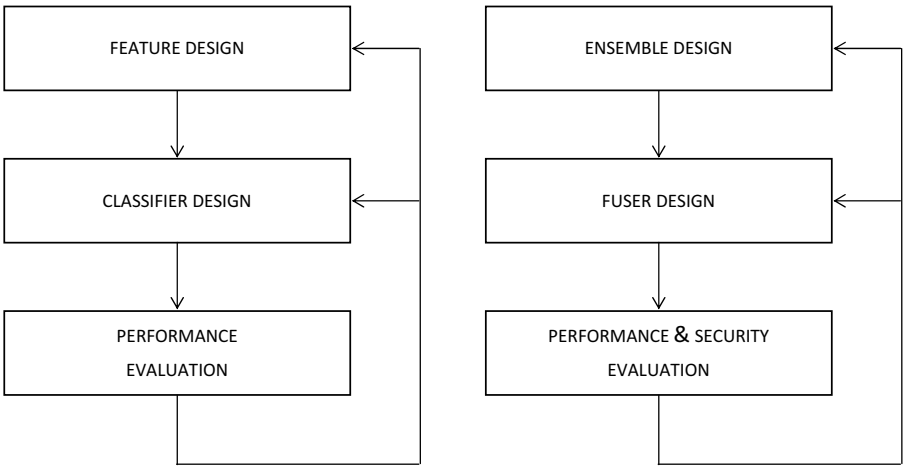


Figure 6.2: Parallel (top left), serial (bottom) and hierarchical (top right) topologies to combine classifier ASV systems and countermeasures.



(a) classifier design process (b) MCS design process

Figure 6.3: Comparison of design processes for a single classifier and for MCS. Diagram adapted from [167, 168]

In general, as stated in [168], no design method guarantees to obtain the "optimal" ensemble for a given fuser or a given application. The best MCS can only be determined by an evaluation procedure. As illustrated in Figure 6.3, a proper evaluation plays a critical role in the design cycle of both single classifiers and MCS. Again, this fact highlight the need to count with standard and well accepted evaluations that also addresses security aspects. From the point of view of the author of this thesis, to address this issue is at the highest priority, and must be in the development of new databases, efforts in different attack and spoofing challenges.

Chapter 7

Countermeasures

This chapter introduces our work on specific approaches to countermeasures to protect ASV systems from spoofing. As stated in Chapter 6, this thesis focuses in the development of countermeasures which are independent of the biometric system they aim to protect.

The three countermeasures presented in this document are selected from a number of approaches developed under the scope of this thesis[1]. The first is a trivial approach based on feature distribution analysis which is effective in detecting artificial signals, the second is based on pairwise distances analysis to detect converted voices and the last uses local binary patterns for generalised attack detection.

Apart from their independence from the ASV system, they share other common characteristics, presented in Section 7.1. Experimental setup and results related to these countermeasures are presented in the next chapter.

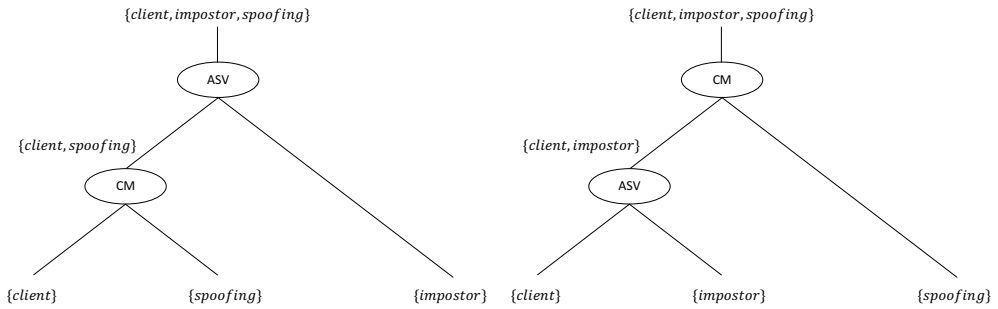
7.1 Generalities

The countermeasures presented in this thesis depend exclusively on the data used for recognition, including training and background data and the same biometric sample used for recognition. This assumptions avoid the use of extra-procedures or interaction with the acquisition system such as challenge-response or to repeat a given sentence e.g. [52].

Countermeasures for voice modality can be divided into speaker-independent and speaker-dependent approaches.

Speaker-independent countermeasures detect electronically generated, manipulated or replayed speech from natural speech. They are strictly related with the liveness detectors defined in Section 3.1.2 and do not consider spoofing

[1]Detailed information about all the countermeasures approaches developed during this thesis can be found in TABULA RASA deliverables



(a) speaker-dependent (b) speaker-independent

Figure 7.1: Decision trees to show how speaker-dependent (Figure 7.1(a)) and speaker-independent countermeasures (Figure 7.1(a)) discriminate the class of interest *client* from the class *impostor* and *spoofing*.

attacks with real speech such as impersonation.

Speaker-dependent countermeasures include liveness detectors helped with speech data from the targeted speaker (client) to perform the spoofing analysis and also (theoretical, not yet developed) mimicry detectors. While the fact the use of client data suggests that speaker-dependent countermeasures may also be useful for speaker discrimination, it is neither the purpose of their design nor subject of evaluation in this thesis.

Figure 7.1 shows, by mean of decision trees, how both approaches can be used to identify the class of interest (client). While speaker-dependent approaches consider the design of purpose-specific approaches against impersonation, speaker-independent approaches relay on the ASV system to detect threats of this kind.

7.1.1 Countermeasure architecture

The countermeasures developed in this thesis have the standard architecture of conventional classifiers, shown in Figure 7.2, which operate in two modes: training (learning) and classification (testing).

The preprocessing module involves operations such as detection of the pattern of interest from the background, noise removal and pattern normalization, among others. In this thesis the speech detection module (Section 2.2.1) is shared by the recognition system and the countermeasure, based on the fact that the information relevant to both speaker recognition and spoofing detection is contained after speech detection.

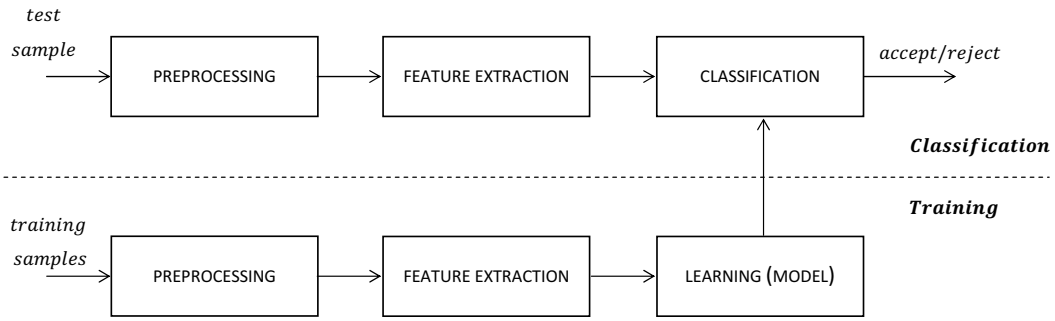


Figure 7.2: Block diagram of the architecture of a generic countermeasure. Illustration adapted from [92].

The feature extraction module finds the appropriate features for representing the input patterns. Features extracted over an interval longer than that used in conventional cepstral analysis are expected to afford a level of protection against spoofing which targets the manipulation of a signal at the single frame, or short duration level.

Higher-level features can be extracted at the multiple frame level, word level, phrase level or even at the utterance level. The features presented in this thesis are utterance-level features i.e. each utterance is represented by a single vector.

In the training mode, a model is generated from a set of features by mean on a learning procedure. For linear classifiers, learning methods can be grouped into two broad classes which are generative and discriminative models [26]. In the classification mode, the trained classifier assigns the input pattern to one of the pattern classes under consideration based on the measured features.

All countermeasures developed in this thesis are based on one-class classifiers (except for a two-class approach included for comparison purposes). Including one-class classification in biometric countermeasures is one of the main contributions in this thesis and thus is described in detail in the next section.

7.1.2 One class classifiers & generalised countermeasures

In Section 6.1 we discuss the motivation to include one-class classifiers in this research. In this section we extend with discussion and describe the classification approach followed in this thesis.

7.1.2.1 Motivation for generalised countermeasures

In comparison to some other biometric modalities, spoofing and countermeasure research in ASV is far less advanced. While attacks from impersonation, replay, speech synthesis and voice conversion are all known, there is a high degree of variation in specific algorithms and there are certainly other forms of attack yet to be identified. Current work in spoofing countermeasures for ASV thus optimistically biases results to known attacks and specific algorithms.

The effect of the use of prior knowledge in the development of countermeasures is illustrated by some examples from previous work. Countermeasures based on the use of phase [198, 200, 53] and prosodic features [145, 56] can be used very successfully to detect voice conversion and speech synthesis attacks. It is likely, however, that they will be overcome by the particular approach to voice conversion investigated in [133] which modifies only the spectral slope of a speech utterance while retaining the original phase and pitch of the original, genuine speech signal.

Another example relates to the average IFDLL proposed in [173]. While this measure is used to successfully detect speech produced by typical HMM-based speech synthesizers, the work in [54] note that it no longer appears to be robust against speech produced by state-of-the-art HMM-based speech synthesizer that include global time variation models [187]. Although the problem is successfully reassessed in [56] by using new measures based on F0 statistics, future speech synthesizers could also make this measure obsolete.

While the true extent of spoofing in the context of ASV is yet to be fully understood, and in any case, there is thus a need for generalised approaches.

7.1.2.2 One-class classifiers

Binary spoofing detectors are generally independent of the biometric system and, opposite to one-class approaches, they are typically trained using both genuine data (negative samples) and examples of spoofed data (positive samples). The main drawback of such an approach is the over fitting to spoofing attacks seen in the training data and thus the lack of generalization to attacks previously unseen [51]. While additional classifiers can be trained and integrated when new attacks are identified, clearly this leads to increased system complexity.

Accordingly in this thesis we have pursued a one-class classification approach to detect spoofing. One-class classifiers differ from two-class and multi-class

classifiers in that only data from one class is used for training, and therefore classifiers are designed to distinguish between the one known class and any other which is unseen during training. Applications of these classifiers are related to anomaly/outlier detection. One class support vector machine (SVM) classifiers are popular, where usually the idea is to minimize the volume of the hypersphere which contains the training data [184].

In this thesis we stress the use of (speech) spoofing data *exclusively* for evaluation purposes. In most of the work on countermeasures reported to date, the spoofing data to learn and to evaluate a given countermeasure is produced in an identical fashion. In this sense, "unseen" attacks include all spoofing signals that present a condition mismatch with respect to the ones generated in the laboratory e.g. mismatch in spoofing technologies or in spoofing system configurations[2] which is the most likely scenario to happen in practice. Therefore, all the spoofing data generated in this thesis is used for evaluation purposes, but not to learn countermeasure classifiers.

We also focus on discriminative approaches; they are usually preferred over the generative ones in related work on countermeasures and they are also preferred in the one-class field when big amount of training data is available, which is our case. In all discriminative (or boundary [184]) one-class classification methods two distinct elements can be identified: a measure for the distance of an object to the target class and a threshold on this distance. New objects are accepted by the description when the distance to the target class is smaller than a given threshold.

Most of the classifiers adopted in this thesis are trivial i.e. distance classifiers, or just the distance between two features in the case of a speaker-dependent countermeasure.

In general, one cannot expect a one-class classifier to have as good performance as a two-class classifier because training samples from two classes provide more information to define the decision boundary than just sampling on one side [184, 209]. Although during evaluation we train binary classifiers e.g. two-class SVM, it is only done for comparison purposes.

The fact that the classifiers used in this thesis are one-class, simple approaches can be interpreted as a stress on the feature extraction phase, which are actually the strong point of the countermeasures presented in this chapter.

[2]Similar examples for other biometrics include photos of different quality of gummy fingers made of different material related with face and fingerprint recognition, respectively

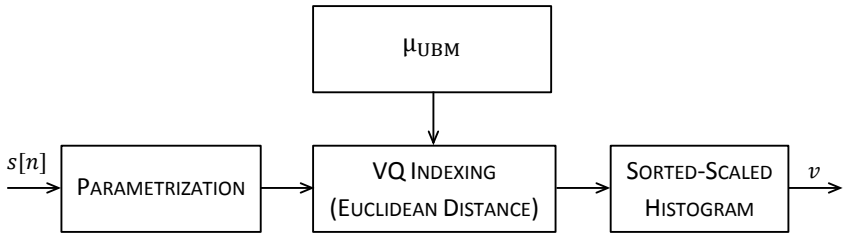


Figure 7.3: Block diagram of repetitive-pattern feature extraction.

7.2 Feature distribution analysis against artificial signals

Some state-of-the-art ASV parameterisations and systems capture and utilise speech characteristics at the utterance level. For example, ASV systems based on GMM supervectors inherently capture speech variability and we thus hypothesize that such systems will be naturally robust to spoofing attacks with artificial signals. This hypothesis is investigated in our experiments reported in Section 5.3. Here we describe a spoofing countermeasure use of longer contexts, which is independent of recognition and is thus applicable to any ASV system.

7.2.1 Repetitive-pattern feature

An implementation of a countermeasure using utterance-level features was developed to protect ASV systems from artificial signals with a repetitive pattern. The following is adapted from the author's own work previously published in [12].

Our approach to utterance-level feature extraction is illustrated in Figure 7.3. Parameters extracted from the input signal are indexed using vector quantization (VQ) and with respect to the means of the universal background model (UBM) components which act as VQ centroids. The histogram of the resulting index vector is reordered based on the occurrence frequencies and the frequencies are scaled with respect to the first component to obtain a single feature vector v.

During the vector quantization step, parameters from a speech signal are expected to be indexed to spread across all or most of the Gaussian components. In contrast, a spoofed signal with repetitive patterns (artifacts, or tone-like artificial signal) will be associated with a smaller number of components. Hence

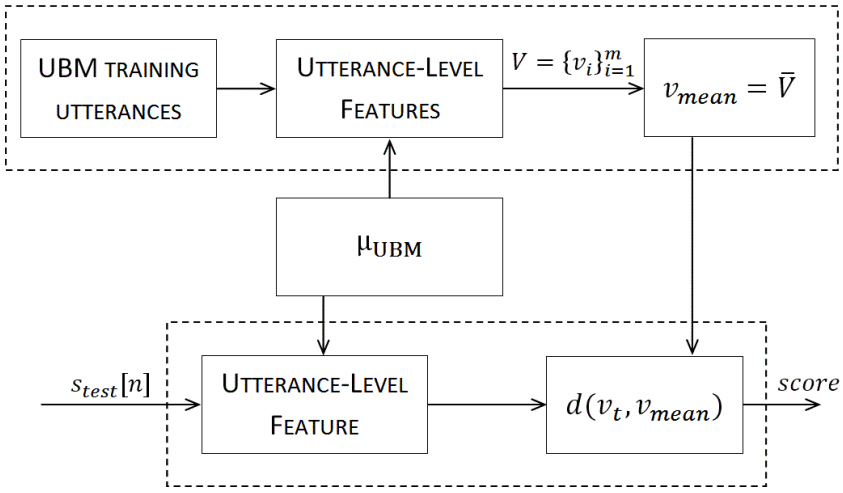


Figure 7.4: Utterance-level features used as an attack detector.

a genuine speech utterance will give rise to a feature vector v with a smooth exponential distribution while a spoofed signal will produce a dirac/delta-like distribution with a smaller number of dominant peaks in the first few coefficients.

7.2.2 Classification and integration

A countermeasure to detect artificial signals using utterance-level features is is implemented by mean of a simple *mean distance classifier*, illustrated in Figure 7.4. This classifier is based on a similarity measure between the test feature vector v_t and a mean feature vector v_{mean} obtained from training examples of genuine speech utterances. For this countermeasure the cosine distance has showed good results and consequently it was chosen as similarity measure. A larger cosine similarity measure indicates a higher probability of spoofing [12].

7.3 Pairwise distances analysis

The work presented in this section is conducted with full prior knowledge of the specific spoofing attack, i.e. the specific algorithm used for the Gaussian dependent filtering (GD-GMM) voice conversion approach originally proposed in [133].

7.3.1 Theoretical framework

This section extends our prior work reported in [6] by making a thorough analysis of the GD-GMM voice conversion system described in Section 3.2.4.2. This system, illustrated in Figure 7.5(b), can be viewed as the implementable version of the not implementable, *ideal* voice conversion system in Figure 7.5(a). While the system in Figure 7.5(a) produces the *optimal* converted speech $\hat{s}[n]$, the system in Figure 7.5(b) produces the suboptimal version $\tilde{s}[n]$.

In this section, the terms *ideal* and *optimal* are defined in a broad sense, as a measure of the level of success of a voice conversion system in applying a particular conversion strategy. In the ideal approach (Figure 7.5(a)), we observe that after front-end processing of the input signal $s[n]$, the features y^k_{asr} are relocated in $\hat{x}^k_{asr}$ according to the following conversion strategy:

$$\hat{x}^k_{asr} = \sum_{i=1}^{M} p(\mu^i_{asr}|y^k_{asr})\mu^i_{asr} \tag{7.1}$$

which is no other than the expectation step in the Expectation Maximization (EM) algorithm [162]. In the ideal system we also assume a perfect signal reconstruction followed by a convenient front-end processing of the targeted ASV system. Therefore, the features $\hat{x}^k_{asr}$ obtained at point *1* are the same features "seen" by the ASV system i.e. point *2* in Figure 7.5(a).

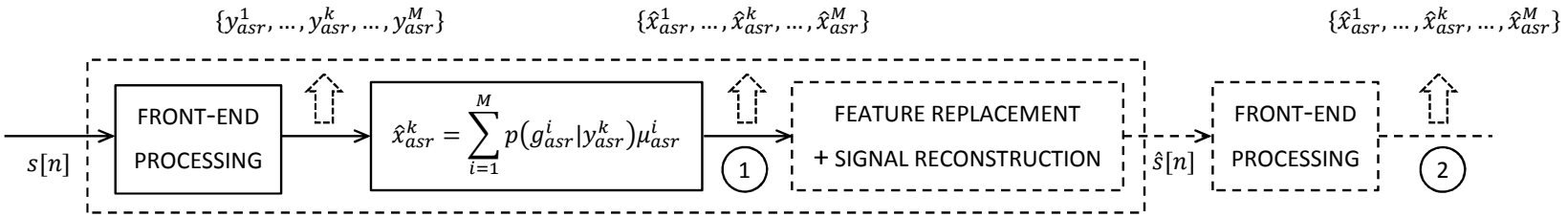
However, the system illustrated in Figure 7.5(a) is not implementable. Matrouf *et al.* [133] observe that the process between points 1 and 2 are unrealisable, since usually the front-end processing is not reversible after feature normalization (the affected modules are thus represented by dashed lines).

In the practical implementation in Figure 7.5(b), the conversion strategy represented by Equation 7.1 is approximated by replacing μ^i_{asr} by the tied models μ^i_{fil}. The resulting Equation is showed as follows:

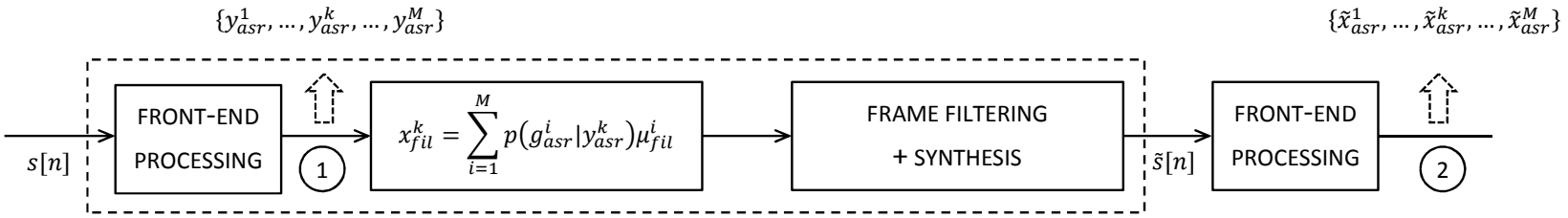
$$x^k_{fil} = \sum_{i=1}^{M} p(\mu^i_{asr}|y^k_{asr})\mu^i_{fil} \tag{7.2}$$

which is identical to Equation 3.6[3]. Finally, x^k_{fil} are used for frame filtering and synthesis. In this case, the same features seen by the ASV system (point *2* in Figure 7.5(b)) correspond to $\tilde{x}^k_{asr}$.

[3]Here we show Equation 7.2 instead of Equation 3.6 so it resembles Matrouf's work

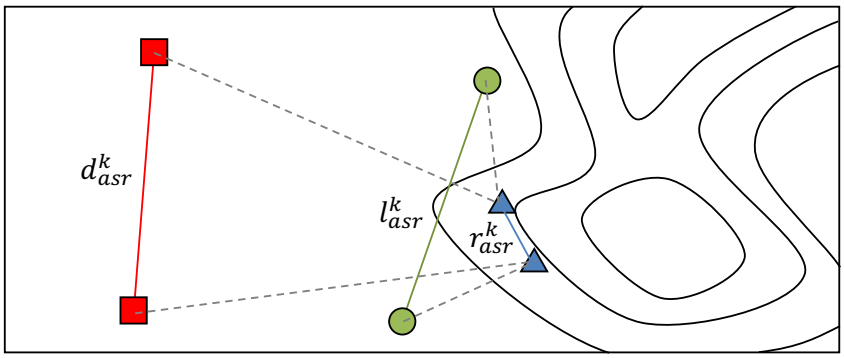


(a) Ideal implementation of voice conversion system

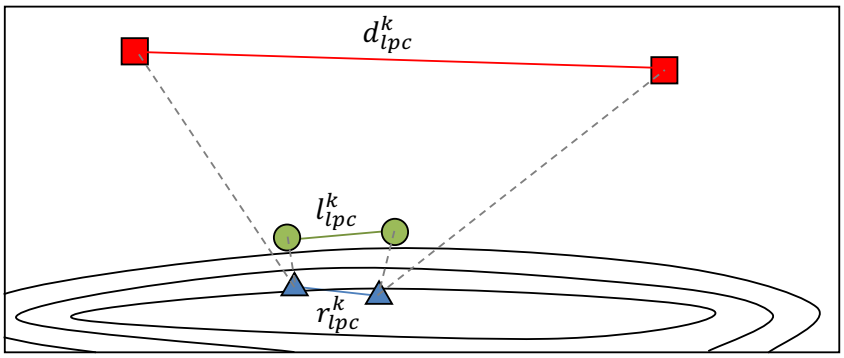


(b) Practical implementation of voice conversion system

Figure 7.5: Comparison between ideal and practical implementations of Matrouf's [133] voice conversion. In ideal voice conversion (larger dashed box in Figure 7.5(a)) feature vectors y^k_{asr} are relocated to $\hat{x}^k_{asr}$. In practical voice conversion (Figure 7.5(b)) μ^i_{asr} is replaced by tied models μ^i_{fil}. Processes between points 1 and 2 generate the k^{th} feature vector in $\tilde{x}^k_{asr}$ instead of $\hat{x}^k_{asr}$



(a) asr space



(b) LPC space

Figure 7.6: An illustration of voice conversion in the ASV system feature space 7.6(a) and the LPC space 7.6(b) showing ideal and real shift of two consecutive vectors. In Figure 7.6(a), conversion aims to relocate original feature vectors y^k_{asr} and y^{k+1}_{asr} (red squares), to $\hat{x}^k_{asr}$ and $\hat{x}^{k+1}_{asr}$ (blue triangles). In practice, due to the necessary use of the implementation in Figure 7.5(b), features are instead relocated as $\tilde{x}^k_{asr}$ and $\tilde{x}^{k+1}_{asr}$ (green circles). Similar analysis for Figure 7.6(b).

7.3.2 PWD feature

We reported a trivial countermeasure against GD-GMM voice conversion in [6] based on the analysis of Equation 7.1. Indeed, the process of mapping a frame onto the space of posteriors and remapping it as a weighted average of features associated with each component causes a shift towards the nearest local maximum. Hence, adjacent frames are expected to be closer to each other after applying such a mapping.

The principal behind our countermeasure exploits the expected shift of consecutive feature frames towards the same, closest local maxima of the like-

lihood function of a particular target model. This principal is illustrated in Figure 7.6(a) for two consecutive feature vectors in two-dimensional space. Under such conditions the relative distance between consecutive feature vectors (red squares) will reduce (blue triangles) i.e. $r^k_{asr} < d^k_{asr}$, whereas the density of features surrounding the local maxima will increase. However, we note that in practice we are interested in the relative distance between $\tilde{x}^k_{asr}$ and $\tilde{x}^{k+1}_{asr}$ i.e. l^k_{asr} in Figure 7.6(a).

We conducted initial experiments to validate this phenomenon. Figure 7.7 shows plots of the $n-1$ consecutive, pairwise distances for n frames of example genuine speech and converted voice signals. As shown in Figure 7.7(b), the differences between genuine speech and converted voice is relatively low in the (normalised) ASR feature space; the two profiles are more or less identical. If the features are not normalised 7.7(a) the differences are more significant as well as for LPCC space 7.7(c) while in LPC space 7.7(d) they are particularly pronounced.

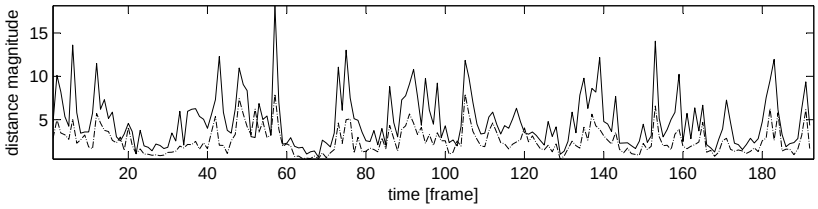
We note that plots 7.7(b), 7.7(c) and 7.7(d) in Figure 7.7 correspond to l^k distances i.e. distances between parameterisations obtained at point 2 in Figure 7.5(b), while plot 7.7(a) is included for comparison purposes.

Related experiments (not reported in this document) have confirmed that $r^k_{asr} < d^k_{asr}$. However, in practice a countermeasure based on this idea is effective as long as r^k is similar to l^k, not the case in the feature space (Figure 7.6(a), illustrated based on Figure 7.7(b)). While the reason of this is still not fully understood, based on comparisons between Figures 7.7(a) and 7.7(b) we hypothesize with a certain degree of confidence that the re-normalization step when re-extracting *asr* features plays an important role on it.

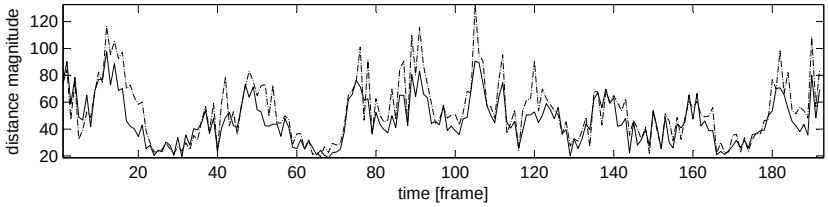
In [6] we reported experiments which showed that distances l^k_{lpc} are not overly dissimilar to r^k_{lpc} and consistently shorter than d^k_{lpc} and that the measure provides a reliable indicator of conversion (Figure 7.6(b)).

For the countermeasure we used alpha coefficients, calculated by using SPRO. We try neither reflection coefficients nor log area ratios. We acknowledge that the alpha coefficients are not the best choice due to their non Gaussian distribution. However, we keep this parametrization simply because we restricted our analysis to the configurations (parameterisations) used in the voice conversion system. We accept the use of such prior knowledge is unrealistic.

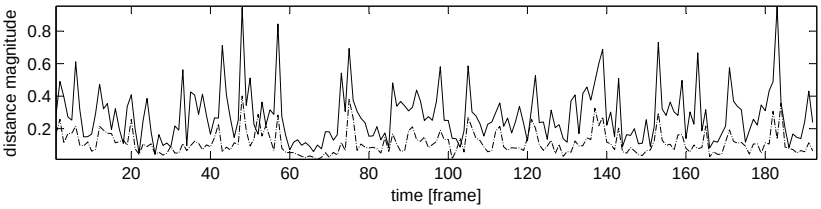
Reassuringly, the original cluster of speech frames represented by their alpha coefficients become more dense as a results of conversion. This fact leaded us to try pairwise distances among every pair of points in the cluster, for



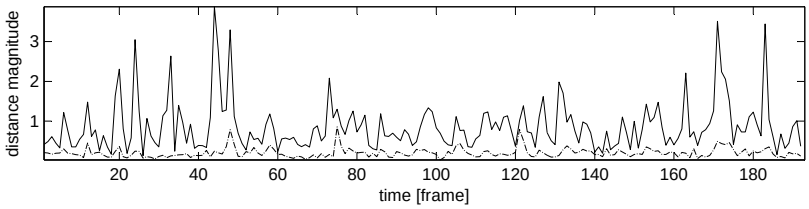
(a) (no-normalised) ASR feature space



(b) (normalised) ASR feature space



(c) LPCC space



(d) LPC space

Figure 7.7: An illustration of the pair-wise distance between consecutive feature vectors for ASR (with and without feature normalization), LPCC and LPC parameterisations. Profiles shown for genuine speech (solid line profiles) and converted voice (dashed profiles).

which pairwise distances between consecutive frames can be seen as a subset. With this new approach, which has the advantage to allow the calculation of a 'softer' distances distribution, we have obtained slightly better results (improvement of 0,3 in the EER%).

Finally, even though we predict increases in cluster density as a result of voice

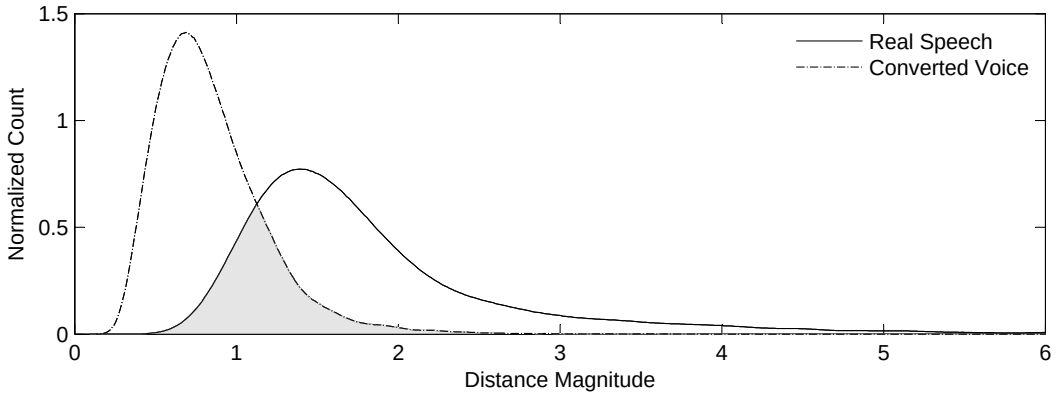


Figure 7.8: An illustration of the LPC distance distribution (Figure 7.7(d)) of 2.5 minutes of real speech (green curve) and its converted version (dashed red curve). The value of the areas behind both curves is normalised to 1. The overlap index (gray area), thus, is a value between 0 and 1.

conversion, initial experiments to observe changes in variance were discouraging and thus the use of variance estimates was not pursued further.

7.3.3 Classification and integration

A block diagram of the integrated ASV system and proposed countermeasure is illustrated in Figure 7.9. The countermeasure is speaker-dependent (as described in Section 7.1 and exploits differences in the distribution of pairwise distances between test data $s[n]$ and that used to train the target model in question.

The similarity measure (score) used in this case is the *overlap coefficient* i.e. the percentage overlap between the two distributions illustrated in Figure 7.8, which is then thresholded to classify $s[n]$ as either genuine speech or converted voice. When the two distributions are normalised, the percentage overlap lies between zero and unity. Lower scores indicate genuine speech whereas higher scores indicate converted voice.

As in other prior work [54, 200], and illustrated in Figure 7.9, the proposed countermeasure is integrated with the ASV system as an independent post processing step. Claimed identities are thus only accepted if a test signal $s[n]$ attains a likelihood higher than the ASV threshold and a countermeasure score lower than its threshold.

Finally, is worth noting that although this countermeasure proved highly ef-

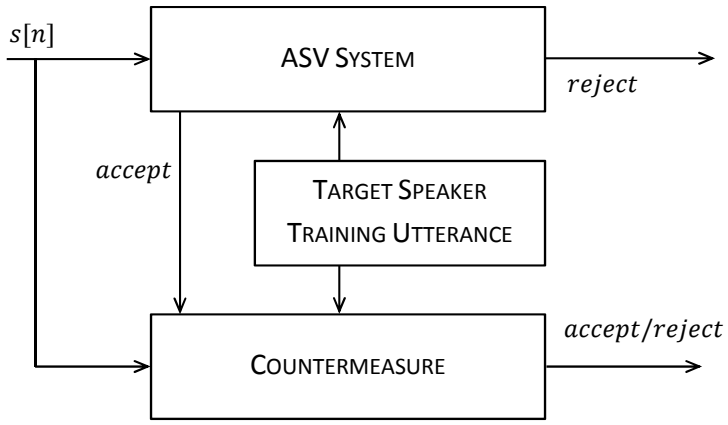


Figure 7.9: A block diagram of the integrated ASV system and proposed countermeasure.

fective, it is conceivably straightforward to overcome: its success is closely dependent on Equation 7.1, which can be modified/improved by the would-be spoofer. This important fact has influenced the research direction of this thesis toward new, generalised countermeasures, presented in next section.

7.4 Local Binary Patterns for generalised countermeasures

For generalised countermeasures we introduce a new feature for spoofing detection based on the analysis of conventional speech parameterisations using standard local binary patterns (LBP). This feature, combined with one-class classifier is, to the best of our knowledge, the first generalised approach to spoofing detection. The following is adapted from the author's own work previously published in [11, 5].

Similar to the PWD countermeasure described in Section 7.3, this work is conducted with full prior knowledge of a single, specific spoofing attack, namely voice conversion, and one more time with the specific approach originally proposed in [133]. However, in contrast to the work described in Section 7.3, no knowledge of alternative attacks (i.e. artificial signals and speech synthesis) was used intentionally during development.

Countermeasure performance is nonetheless assessed in the case of all four spoofing attacks. The goal is then to show that this new approach has potential to detect previously unseen attacks for which the countermeasure is not

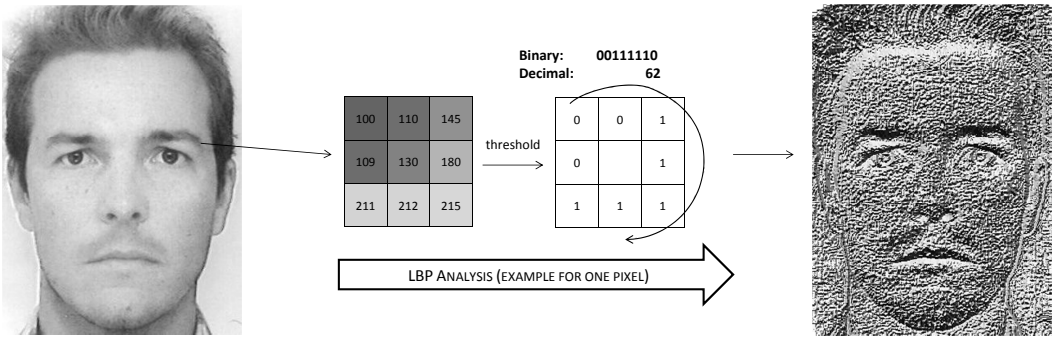


Figure 7.10: Illustration of the original LBP operator for face images. Each pixel of the original image (left) can have one of 256 possible intensity values (0-255). The figure shows an example of a pixel and its neighbouring pixels. After LBP analysis, the intensity value of each pixel is replaced by its LBP code (the value 130 is replaced by 62 for the example in the figure).

optimised. This condition is the most challenging considered thus far, and the most representative of the practical scenario where the nature of the spoofing attacks can never be known.

7.4.1 LBP features

The new countermeasure is based on the hypothesis that modifications made through spoofing disturb the natural, dynamic spectro-temporal "texture" of genuine speech. Motivated by the fact that computer vision techniques were already successfully applied in the speech field [170], we have investigated the application of a standard texture analysis approach, known as Local Binary Patterns [147], to a 2-dimensional "image" of a speech utterance, where here the image is a linear-scaled cepstrogram appended with dynamic features.

The local binary pattern (LBP) is a non-parametric operator which describes the local spatial structure of an image. The original Local Binary Pattern (LBP) operator first introduced by *Ojala et al.* [146] is a 3x3 kernel which assigns a binary code to each pixel in an image according to the comparison of its intensity value to that of its eight surrounding pixels.

The procedure is illustrated in Figure 7.10. At a given pixel position (x_c, y_c), a binary value of 0 is assigned when the intensity of neighbouring pixels is lower, whereas a value of 1 is assigned when neighbouring pixels are of higher or equal intensity. Each pixel is thus assigned one of $2^8 = 256$ binary patterns.

The decimal form of the resulting 8-bit word (LBP code) can be expressed as

follows:

$$LBP(x_c, y_c) = \sum_{n=0}^{7} s(i_n - i_c)2^n \tag{7.3}$$

where i_c corresponds to the grey value of the center pixel (x_c, y_c), i_n to the grey values of the 8 surrounding pixels, and function $s(x)$ is defined as:

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{7.4}$$

LBP has become very popular in pattern recognition due to its texture discriminative property and its very low computational cost. Main applications include face detection [96], face recognition [212, 3], image retrieval [182], motion detection [89] or visual inspection [188]. Figure 7.10 shows a face image before and after LBP analysis.

In this work we reduce the number of possible patterns according to the standard Uniform LBP approach described in [147]. Uniform LBPs are the subset of 58 patterns which contain at most two bitwise transitions from 0 to 1 or 1 to 0 when the bit pattern is traversed in circular fashion. As an example, the subset includes patterns 00000001 and 00111100 but not 00110001.

As reported by [147], most patterns are naturally uniform and empirical evidence suggests that their use in many image recognition applications leads to better performance than the full set of uniform and non-uniform patterns. We observed similar findings in our work and thus pixels corresponding to any of the 198 non-uniform patterns are simply ignored.

The procedure to obtain the anti-spoofing LBP feature is illustrated in Figure 7.11. LBPs are determined for each pixel in the linear-scaled cepstrogram thus resulting in a new matrix of reduced dynamic range, here referred to as a *textrogram*. The textrogram captures short-time feature motion beyond that in conventional dynamic parameterisations. The LBP-based countermeasure is based on concatenated histograms formed from the pixel values across each row in the textrogram. The histograms are individually normalised and their resulting bin values are stacked vertically to obtain a new vector in the same manner as GMM mean-vectors are stacked to form supervectors.

Figure 7.11: Application of uniform LBP analysis to obtain a textrogram from a matrix formed from the concatenation of conventional feature vectors. Non-uniform patterns (blank cells in textrogram) are discarded and the resulting feature used for spoofing detection is formed from the concatenation of normalised histograms of the remaining uniform codes in each row. Figure reproduced from [11].

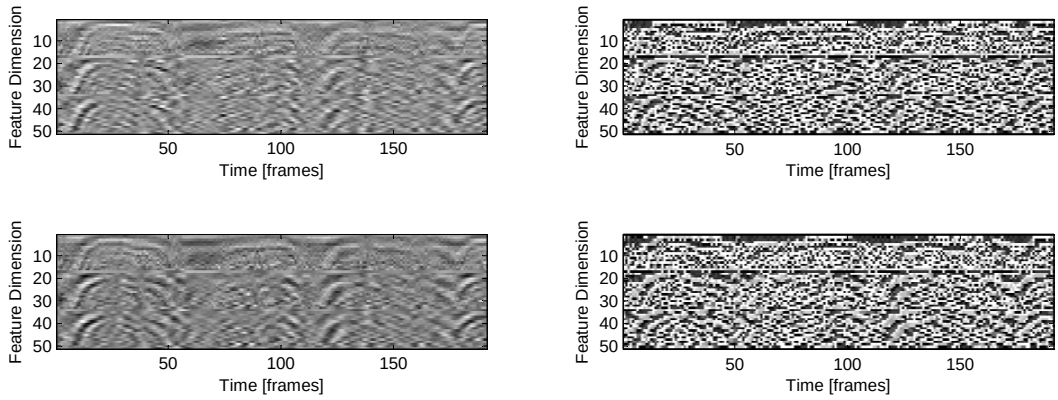


Figure 7.12: On the left: Example of concatenated feature vectors extracted from 193 consecutive speech frames (approximately 2 seconds of continue speech) of real speech (above) and its converted version (below). On the right: uniform LBP operator applied to feature vectors. Note that each 'image' is comes from approx. 2.5 min of speech (around 10000 frames). Figure reproduced from [11].

The division of the textrogram (or equivalent in image recognition problems) is also standard practice [3] and serves to provide a greater level of granularity than would be provided with only a single histogram corresponding to the full textrogram.

Example cepstrograms (left) and textograms (right) are illustrated in Figure 7.12 for both genuine speech (top) and a spoofed attack through voice conversion (bottom). While a certain level of smoothing is detectable in the cepstrograms, differences in the textrograms are more pronounced (although not immediately obvious by eye) and point to the potential of the new approach to detect spoofing.

7.4.2 Classification and integration

The countermeasure is integrated into a full ASV system as an independent classifier in equivalent fashion to the Figure 7.9. Figure 7.13 illustrates the countermeasure in stand-alone operation, in which the LBP-based feature is used as an input of an one-class classifier. As mentioned before, the classifier used for the spectral texture analysis countermeasure aims to improve generality to previously unseen spoofing attacks.

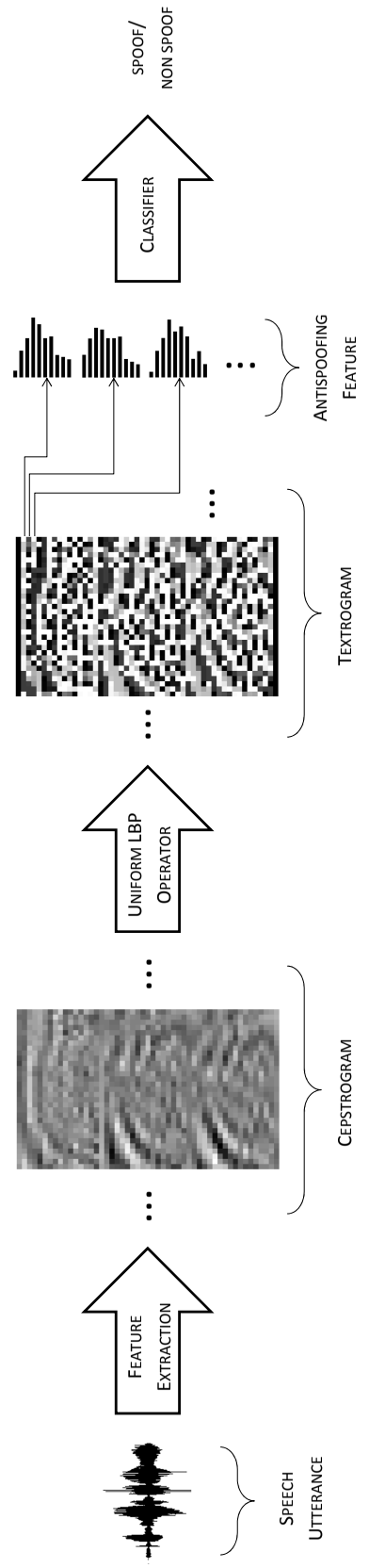


Figure 7.13: LBP-based countermeasure in testing mode (merge of Figures 7.2, 7.13 and 7.11). The figure shows the application of uniform LBP analysis to a cepstrogram to obtain the so-called textrogram and the resulting feature as an input of a (in this thesis one-class) classifier. Figure reproduced from [5].

Two countermeasures are proposed based on LBP features and two different one-class classifiers, a speaker-dependent [11] and an speaker-independent [5] approach. For the speaker-dependent case, LBP-based features are calculated for the test sample and that used for training client model in the ASV system. The two resulting feature vectors are compared using histogram intersection and the resulting score is thresholded to classify the test signal as genuine speech or a spoofing attack. For the speaker-independent approach, the LBP feature calculated from a test sample is tested against model resulting from a one-class SVM learning [11, 5] (we note that in this case the amount of training data is considerable). Experimental work on both countermeasures is reported in next chapter.

Chapter 8

Evaluation: Countermeasures & integration

Baseline and spoofing related experiments were reported in Chapter 5. Since there are numerous, different approaches to speaker recognition in the literature, both were studied with a range of systems: a standard Gaussian mixture model with universal background model (GMM-UBM), a GMM supervector linear kernel (GSL) system, a GMM supervector linear kernel system with nuisance attribute projection (GSL-NAP), a factor analysis (FA) system, a GSL system with FA supervectors (GSL-FA) and an i-vector system with PLDA postprocessing.

We also considered three high-effort spoofing attacks i.e. artificial signals, voice conversion and synthesized speech and also one example of a low-effort attack with white noise which is not addressed in this chapter.

In the following we evaluate the performance of the three countermeasures introduced in previous chapter. Evaluations include performance assessment in stand-alone operation and also in joint-operation with the baseline ASV systems previously defined for this thesis. Specifications and countermeasures setup are described in Section 8.1, results are presented in Section 8.2 and are discussed in Section 8.2.4

8.1 Specifications for countermeasures

Three different approaches have been considered. They are the repetitive pattern detection (RPD) approach developed for artificial signals, pair-wise distances analysis (PWD) for protection against attack with voice conversion and local binary patterns (LBP) based countermeasure for generalised protection.

8.1.1 Countermeasures setup

The countermeasures presented in this section follow the guidelines established in Section 7.1. They are independent of the biometric system they aim to protect, they use the same data (i.e. background data, training samples) used in the targeted ASV systems and they can be speaker-dependent or speaker-independent, among other characteristics.

8.1.1.1 RPD-based countermeasure

The countermeasure basically performs an analysis of feature distribution of the targeted ASV system. Consequently, the front-end processing and the UBM model in Figure 7.3 as well as the UBM training utterances in Figure 7.4 are the ones defined in Section 5.1.1 for the ASV systems setup.

8.1.1.2 PWD-based countermeasure

The countermeasure operates on the same 19^{th} order LPC vectors recalculated from a time domain signal $s[n]$ in Figure 7.9. Frame blocking is the same as for ASV systems and voice conversion (although different frame lengths do provide similar results). We take into account only those frames determined to contain voiced speech. Voiced speech was detected using the robust algorithm for pitch tracking (RAPT) [183] in the VOICEBOX toolkit[1] with a default configuration.

8.1.1.3 LBP-based countermeasure

LBP analysis is applied to cepstrograms composed of 51 coefficients: 16 LFCCs and energy plus their corresponding delta and delta-delta coefficients. Frame blocking is the same as for ASV systems (although different frame lengths do provide similar results). We take into account only those frames determined to contain speech, i.e. those also used for ASV.

We performed experiments with $LBP_{4,1}$, $LBP_{8,1}$, $LBP_{8,2}$, and $LBP_{16,2}$ operators and their uniform versions using the publicly available LBP Matlab implementation from the University of Oulu[2]. Our best results were obtained with a $LBP^{u2}_{8,1}$ operator considering only the 58 possible uniform patterns.

[1] `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`
[2] `http://www.cse.oulu.fi/CMV/Downloads/LBPMatlab`

Histograms are created for all but the first and last rows of the textrogram, thereby obtaining a $58 \times (51 - 2) = 2842$ length feature vector.

We assess three different classifiers. In all cases attacks with speech synthesis and artificial signals represent the universe of unknown attacks. For the first two, one-class approaches, only converted voice was used for optimisation (tuning as opposed to training).

The first classifier is one-class[3], speaker-dependent approach whereby scores correspond to the comparison of the LBP feature vector extracted from the input utterance to that of the target client (from the ASV training dataset) using a histogram intersection kernel.

The second classifier is a one-class, speaker-independent SVM approach where scores correspond to the comparison of the input utterance to the set of LBP features extracted from all utterances in the NIST'04 and NIST'08 datasets (approximately 8000 utterances).

The third classifier is a two-class SVM where each of the two models are trained on the same genuine speech as the second classifier and the 9892 converted voice utterances in the development set respectively.

All SVM classifiers are implemented using the LIBSVM[4] library [44] and are tuned using only genuine speech or converted voices in the development set.

8.1.2 Protocols & metrics

The work in this thesis related to countermeasure evaluations has been conducted according to some general guidelines. First, each proposed countermeasure is evaluated in stand-alone operation against all three spoofing attacks and with the same spoofing database and protocols defined in Section 5.2. Therefore, each proposed countermeasure is evaluated together with each of the six defined ASV systems.

In addition to system independent countermeasure assessments, results are presented through a set of three DET plots containing four profiles each (see Section 3.3.1). Together they represent system performance under spoofing attacks for three different countermeasure operating points defined according to the false fake rejection (FFR) rate (FFR=1%, FFR=5% and FFR=10%).

However, in view of the number of different experimental combinations for the

[3]Only real speech is used for modeling.

[4]`http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

voice modality, only a selection of DET plots are presented. Other results are presented using tables which aim to provide a more concise summary for an FRR of 10% in each case (note that the operative point set to 10% provides additional information on spoofing assessment of that of the Tables 5.3).

Finally, int this evaluation (Section 3.3.1) the countermeasure evaluated standalone is subjected to two kind of errors denoted False Living Rate (FLR) and False Fake Rate (FFR), which are similar to FAR and FRR, respectively. The point at which FLR=FFR is called Average Classification Error (ACE), which is similar to the Equal Error Rate (EER).

8.2 Results

Experimental work related to the repetitive pattern detection (RPD), the pair-wise distances analysis (PWD) and the local binary patterns (LBP) based countermeasures are reported in Section 8.2.1, Section 8.2.2 and Section 8.2.3, respectively. While in the previous chapter baseline results and vulnerabilities to spoofing are assessed for ASV systems with and without score normalisation, evaluation of countermeasures are presented only for the latter.

8.2.1 Repetitive pattern detector

The countermeasure based on the distribution analysis of ASV features was designed to detect repetitive patterns which make spoofing attacks using tone-like, artificial signals such a threat. Significant improvements over the baseline performance were therefore expected in this case. A DET curve which aims to assess the performance of the countermeasure independently from the ASV system is shown in Figure 8.1. As expected the countermeasure is extremely effective in the case of artificial signals and detects all spoofing attacks. In contrast, however, the Average Classification Error (ACE), where the FLR and the FFR have the same value, is almost 40% for voice conversion and 20% for synthesized speech.

System dependent results for the GMM-UBM baseline system are illustrated in Figure 8.2 for NIST evaluation datasets and spoofing attacks with artificial signals. Together the three plots illustrate the efficacy of the higher-level features countermeasure for the three different FFR operating points.

In all cases, almost all of the attacks are detected. Furthermore, the degra-

dation in the performance of the baseline is minimal for all but the lowest FRRs. A summary of results for all six ASV systems and the three different operating points is presented in Table 8.1(a). Each cell contains four values of FAR for (i) the baseline, (ii) the baseline with countermeasures, (iii) the baseline under spoofing attack and (iv) the baseline under spoofing attack, but with countermeasures.

The countermeasure is seen to perform best when the countermeasure is tuned to obtain an FFR of 1%; the FARs for (ii) and (iii) are lowest. While the countermeasure is nonetheless effective in detecting all attacks for all system/FFR combinations, increases in FAR are high in some cases, i.e. when the FFR of the countermeasure is tuned to 10%.

Slight improvements were also observed in the case of synthesized speech. Table 8.1(c) shows that, for a fixed FFR of 5%, the FAR for GMM-UBM and FA systems drops from 83% to 53% and from 61% to 43% respectively. Table 8.1(c) also shows that there is no improvement in performance for GSL-based systems since, as discussed in Section 5.3.2, they are already somewhat robust to synthesized speech. Finally, as illustrated in Table 8.1(b), RPD features are wholly ineffective in the case of voice conversion.

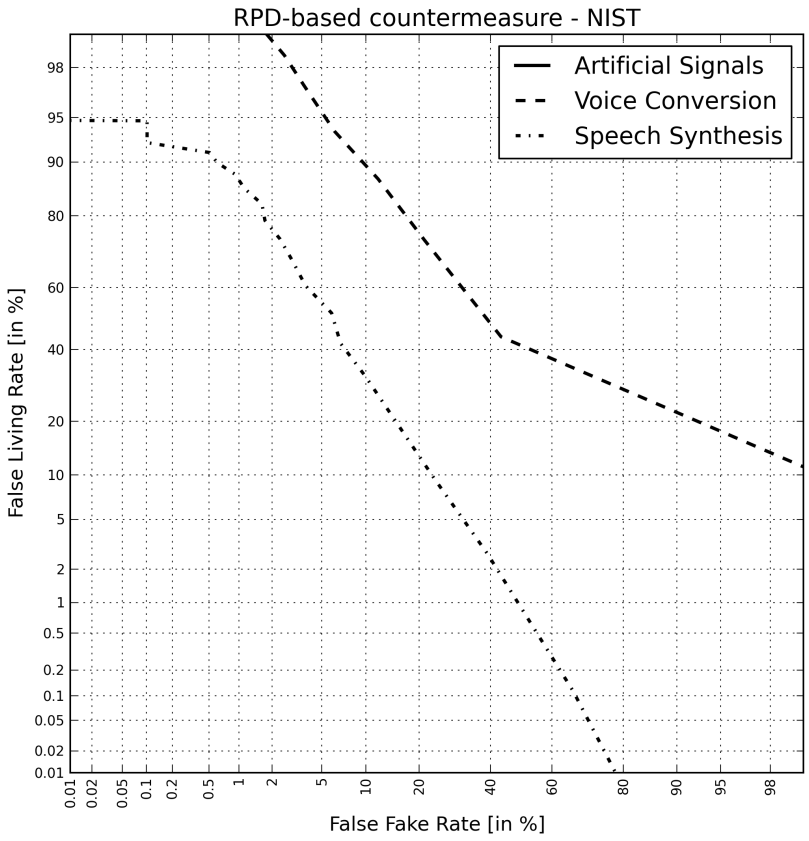


Figure 8.1: DET profiles for the higher-level features countermeasure assessed independently from the ASV system and for the mobile/telephony scenario (Note: the ACE for Artificial signals is equal to 0).

Figure 8.2: DET profiles for the baseline GMM-UBM system with and without the proposed higher-level features countermeasure and for the mobile/telephony scenario. The three figures represent system performance with and without artificial signal spoofing attacks and for the three different countermeasure operating points (FFR=1%, FFR=5% and FFR=10%).

(a) Artificial signals

ASV System	FFR		
	1%	5%	10%
GMM-UBM	7-7-91-0	7-10-91-0	7-29-91-0
GSL	6-7-2-0	6-14-2-0	6-31-2-0
GSL-NAP	3-4-4-0	3-9-4-0	3-36-4-0
GSL-FA	2-2-1-0	2-7-1-0	2-52-1-0
FA	1-1-71-0	1-3-71-0	1-10-71-0
IV-PLDA	0.2-0.5-1-0	0.2-2-1-0	0.2-7-1-0

(b) Voice conversion

ASV System	FFR		
	1%	5%	10%
GMM-UBM	7-7-78-79	7-10-78-82	7-29-78-87
GSL	6-7-89-89	6-14-89-91	6-31-89-89
GSL-NAP	3-4-83-85	3-9-83-88	3-36-83-89
GSL-FA	2-2-82-83	2-7-82-88	2-52-82-90
FA	1-1-54-56	1-3-54-67	1-10-54-79
IV-PLDA	0.2-0.5-55-58	0.2-2-55-72	0.2-7-55-82

(c) Speech synthesis

ASV System	FFR		
	1%	5%	10%
GMM-UBM	7-8-83-76	7-11-83-53	7-36-83-37
GSL	6-7-31-32	6-14-31-33	6-40-31-35
GSL-NAP	3-4-30-30	3-10-30-32	3-40-30-35
GSL-FA	2-2-18-19	2-7-18-25	2-52-18-37
FA	1-1-61-55	1-3-61-43	1-10-61-32
IV-PLDA	0.2-0.5-12-15	0.2-2-12-22	0.2-7-12-35

Table 8.1: False acceptance rate (FAR) scores (%) for a fixed false rejection rate (FRR) of 10% for each of the six speaker verification systems in the mobile/telephony scenario under attack with artificial signals, voice conversion and synthesized speech. Each cell shows the FAR according to the following conditions: (baseline)-(baseline + RPD)-(baseline under attack)-(baseline + RPD under attack).

8.2.2 Pair-wise distances analysis

This work reports experimental results on new countermeasure which exploits the reduction in pair-wise distances between consecutive feature vectors.

The pair-wise distance countermeasure was designed specifically to address the problem of voice conversion. Moreover, the countermeasure was optimized for the mobile/telephony scenario. Significant improvements over the baseline performance were therefore expected in this case.

A DET curve which aims to assess the performance of the countermeasure independently from the ASV system is shown in Figure 8.3. As expected the countermeasure is very effective in the case of voice conversion and detects most of the spoofing attacks (ACE of 2.5%). In contrast, however, the ACE is 35% for artificial signals and 10% for synthesized speech.

An example of DET profiles for the baseline GMM-UBM system with and without the proposed PWD based countermeasure is shown in Figure 8.4. We observe that, for a fixed FFR of 1%, the countermeasure offers satisfactory protection under spoofing attacks with voice conversion, while the degradation in the performance of the baseline is minimal. For fixed FFRs of 5% and 10%, the degradation in baseline performance is non-negligible, even if almost all spoofing attacks are successfully detected.

A summary of results for all six ASV systems and the three different operating points is presented in Table 8.2(b). The countermeasure is seen to perform best when the countermeasure is tuned to obtain an FFR of 1%; the FARs for (ii) and (iii) are lowest. While the countermeasure is nonetheless effective in detecting all attacks for all system/FFR combinations, increases in FAR are high in some cases, i.e. when the FFR of the countermeasure is tuned to 10%.

Slight improvements were also observed in the case of synthesized speech. Table 8.2(c) shows that, for a fixed FFR of 5%, the FAR for GMM-UBM and FA systems drops from 83% to 45% and from 61% to 33% respectively. Table 8.1(c) also shows that there is no improvement in performance for GSL-based systems since, as discussed in Section 5.3, they are already somewhat robust to synthesized speech. Finally, as illustrated in Table 8.2(a), the PWD countermeasure is wholly ineffective in the case of artificial signals.

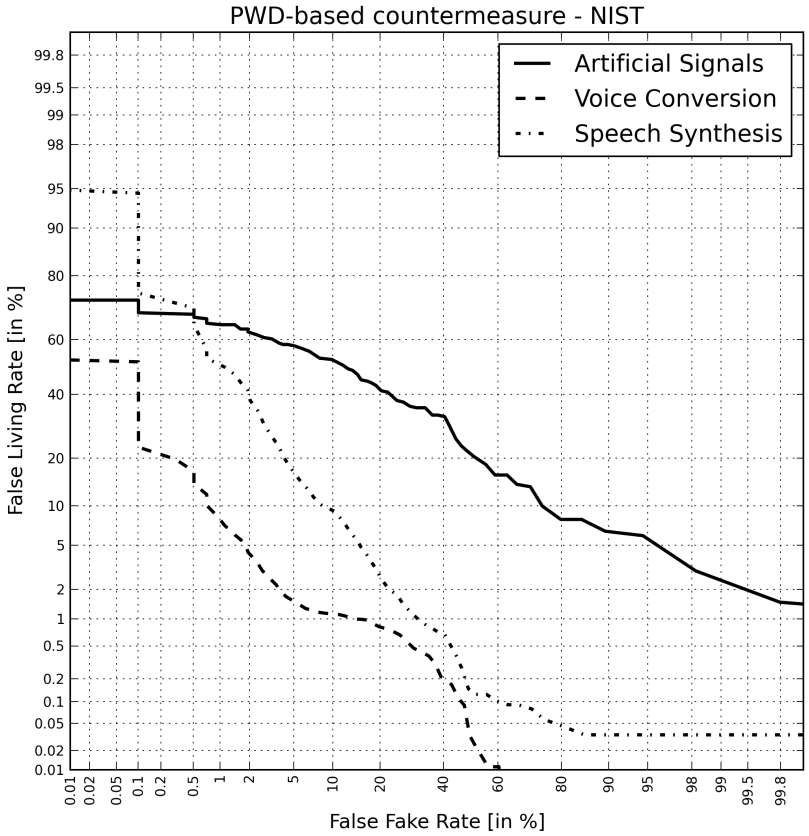


Figure 8.3: DET profiles for the PWD-based countermeasure assessed independently from the ASV system and for the mobile/telephony scenario.

Figure 8.4: DET profiles for the baseline GMM-UBM system with and without the proposed PWD based countermeasure and for the mobile/telephony scenario. The three figures represent system performance with and without attacks with voice conversion and for the three different countermeasure operating points (FFR=1%, FFR=5% and FFR=10%).

(a) Artificial signals

ASV System	FFR		
	1%	5%	10%
GMM-UBM	7-8-91-58	7-10-91-53	7-29-91-49
GSL	6-7-2-2	6-13-2-5	6-28-2-16
GSL-NAP	3-4-4-3	3-7-4-5	3-36-4-35
GSL-FA	2-2-1-1	2-4-1-2	2-41-1-31
FA	1-1-71-46	1-2-71-42	1-9-71-40
IV-PLDA	0.2-0.5-1-1	0.2-1-1-4	0.2-6-1-18

(b) Voice conversion

ASV System	FFR		
	1%	5%	10%
GMM-UBM	7-8-78-6	7-10-78-1	7-29-78-0
GSL	6-7-89-7	6-13-89-1	6-28-89-0
GSL-NAP	3-4-83-6	3-7-83-1	3-36-83-0
GSL-FA	2-2-82-7	2-4-82-1	2-41-82-0
FA	1-1-54-4	1-2-54-0	1-9-54-0
IV-PLDA	0.2-0.5-55-3	0.2-1-55-0	0.2-6-55-0

(c) Speech synthesis

ASV System	FFR		
	1%	5%	10%
GMM-UBM	7-8-83-45	7-10-83-19	7-29-83-10
GSL	6-7-31-20	6-13-31-12	6-28-31-7
GSL-NAP	3-4-30-18	3-7-30-9	3-36-30-8
GSL-FA	2-2-18-11	2-4-18-7	2-41-18-8
FA	1-1-61-33	1-2-61-14	1-9-61-8
IV-PLDA	0.2-0.5-12-11	0.2-1-12-10	0.2-6-12-8

Table 8.2: False acceptance rate (FAR) scores (%) for a fixed false rejection rate (FRR) of 10% for each of the six speaker verification systems in the mobile/telephony scenario under attack with artificial signals, voice conversion and synthesized speech. Each cell shows the FAR according to the following conditions: (baseline)-(baseline + PWD)-(baseline under attack)-(baseline + PWD under attack).

8.2.3 LBP-based countermeasure

We introduced a new feature for spoofing detection based on the analysis of conventional speech parameterisation using standard local binary patterns (LBP). This feature, together with what is, to the best of our knowledge, the first one-class classification approach to spoofing detection, results in a generalised spoofing countermeasure for ASV.

Results for the new LBP-based countermeasure and each of the three different classifiers are illustrated in Table 8.3. For the one-class SVM classifier, we obtained our best results with a radial kernel basis function, while a linear kernel gave better results for the two-class classifier.

As expected, compared to the one-class classifiers, the two-class classifier offers the best performance for the condition on which it is optimised (voice conversion). Here the ACE is 0%. However, for the two spoofing attacks not seen during optimisation, performance is poor. Since the binary SVM classifier is not designed to manage "outliers" it is perhaps not surprising in this case that ACEs increase rather than decrease.

While the one-class classifiers do not perform as well as the two-class classifier for voice conversion spoofing attacks, ACEs of 8% and 5% are only marginally higher than the baseline ACE of 3%. More importantly, the one-class classifiers are seen to generalise well to synthesised speech and artificial signals. Here the ACEs are all less than 1%.

Table 8.3 shows that the best overall performance is obtained with the one-class SVM classifier. However, in this section we focus on the countermeasure with histogram intersection kernel classifier, which is also the chosen for being integrated to the ASV systems. A detection error trade-off (DET) profile which shows the performance of the countermeasure independently from ASV is illustrated in Figure 8.5.

In the following we present an evaluation example that involves GMM-UBM ASV systems, attacks with speech synthesis and the countermeasure with histogram intersection kernel classifier. For further results involving other systems e.g. IV-PLDA and one-class SVM classifiers readers are referred to [5].

DET profiles for the baseline GMM-UBM system with and without the proposed LBP based countermeasure for attacks with speech synthesis is shown in Figure 8.6. A summary of system-dependent results for spoofing attacks with synthesized speech and the LBP countermeasure is also presented in Table 8.4(a).

Classifier Attack	1-class spk-dep	1-class SVM	2-class SVM
Voice Conversion	8	5	0
Speech Synthesis	1	0.1	56
Artificial Signals	0	0	25

Table 8.3: Countermeasure performance in terms of ACE (%) for the three different classifiers and three different spoofing attacks.

Countermeasure performance for spoofing attacks with converted voices is very satisfactory and almost as good as that for the PWD countermeasure (ERR equal to 6.2% for LBP versus 2.5% for PWD). On the other hand, without any specific training data, nor any further optimisation, the new countermeasure is able to detect all attacks performed with artificial signals and almost all attacks with speech synthesis.

A summary of countermeasure performance for all six ASV systems for artificial signals spoofing attacks, voice conversion and synthesized speech is presented in Tables 8.4(a), 8.4(b) and 8.4(c) respectively. Together they illustrate the reliable detection performance for all considered attacks, even though knowledge of only one specific attack is used for optimisation. For voice conversion and a fixed FFR of 1%, the FAR decreases by 50% relative.

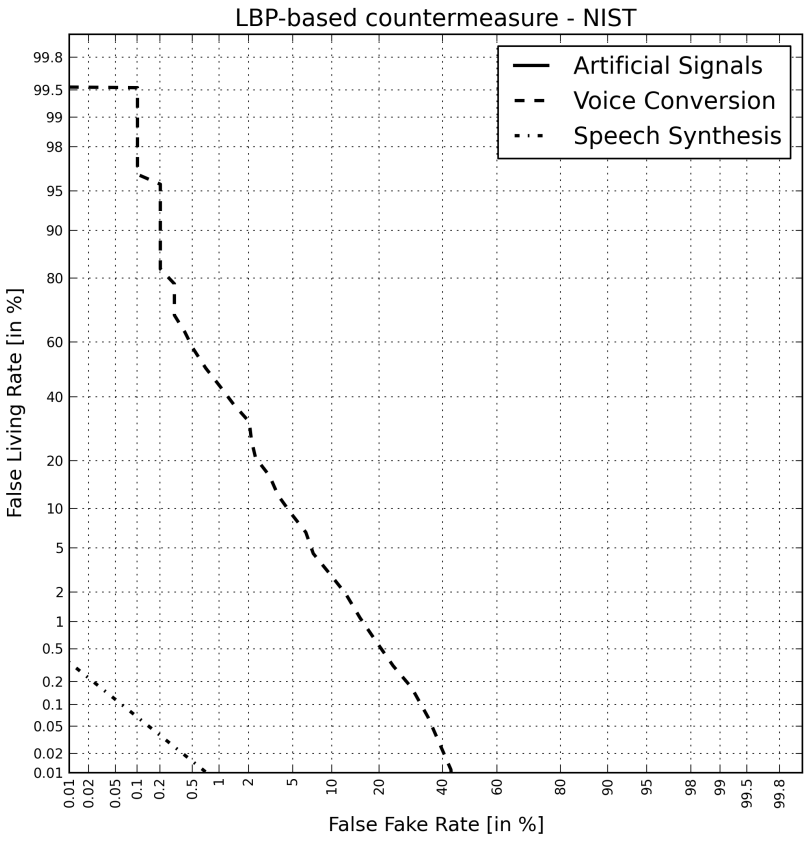


Figure 8.5: DET profiles for the local binary pattern countermeasure assessed independently from the ASV system and for the mobile/telephony scenario (Note: the ACE for Artificial signals is equal to 0).

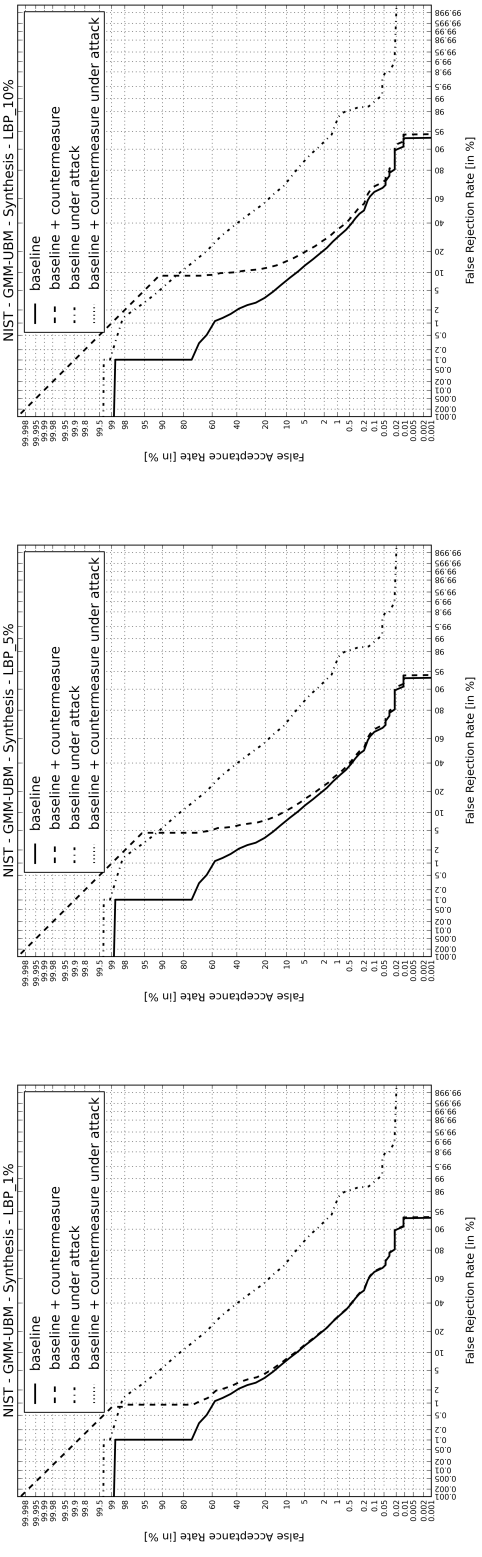


Figure 8.6: DET profiles for the baseline GMM-UBM system with and without the proposed LBP countermeasure and for the mobile/telephony scenario. The three figures represent system performance with and without speech synthesis spoofing attacks and for the three different countermeasure operating points (FFR=1%, FFR=5% and FFR=10%).

(a) Artificial signals

ASV System	FFR		
	1%	5%	10%
GMM-UBM	7-8-91-0	7-12-91-0	7-49-91-0
SGL	6-7-2-0	6-15-2-0	6-61-2-0
SGL-NAP	3-4-4-0	3-12-4-0	3-59-4-0
SGL-FA	2-3-1-0	2-9-1-0	2-68-1-0
FA	1-1-71-0	1-3-71-0	1-16-71-0
IV-PLDA	0.2-0.4-1-0	0.2-1-1-0	0.2-6-1-0

(b) Voice conversion

ASV System	FFR		
	1%	5%	10%
GMM-UBM	7-8-78-38	7-12-78-9	7-49-78-3
SGL	6-7-89-44	6-15-89-10	6-61-89-3
SGL-NAP	3-4-83-42	3-12-83-10	3-59-83-3
SGL-FA	2-3-82-41	2-9-82-10	2-68-82-3
FA	1-1-54-28	1-3-54-8	1-16-54-3
IV-PLDA	0.2-0.4-55-30	0.2-1-55-6	0.2-6-55-5

(c) Synthesized speech

ASV System	FFR		
	1%	5%	10%
GMM-UBM	7-8-83-0	7-12-83-0	7-49-83-0
SGL	6-7-31-0	6-15-31-0	6-61-31-0
SGL-NAP	3-4-30-0	3-12-30-0	3-59-30-0
SGL-FA	2-3-18-0	2-9-18-0	2-68-18-0
FA	1-1-61-0	1-3-61-0	1-16-61-0
IV-PLDA	0.2-0.4-12-0	0.2-1-12-0	0.2-6-12-0

Table 8.4: False acceptance rate (FAR) scores (%) for a fixed false rejection rate (FRR) of 10% for each of the six speaker verification systems in the mobile/telephony scenario under attack with artificial signals, voice conversion and synthesized speech. Each cell shows the FAR according to the following conditions: (baseline)-(baseline + LBP)-(baseline under attack)-(baseline + LBP under attack).

Countermeasure		Attack		
Feature + Classifier	Speaker	VC	SS	AS
RPD + 1-class MDC [12]	independent	40	20	**0**
PWD + 1-class OIC [6]	dependent	**3**	10	35
LBP + 1-class HIK [11]	dependent	**8**	1	0
LBP + 1-class SVM [5]	independent	**5**	0.1	0
LBP + 2-class SVM [5]	independent	**0**	56	25

Table 8.5: Countermeasure performance in terms of ACE (%) in the mobile/telephony scenario for each of the five countermeasures presented in this thesis and attacks with voice conversion (VC), speech synthesis (SS) and artificial signals (AS). Values in bold correspond to the attacks for which the countermeasure was tuned/optimized or trained

8.2.4 Discussion

Table 8.5 summarizes the results related to the performance of the different countermeasures presented in this thesis. They are repetitive pattern detector (RPD) features combined with mean distance classifiers (MDC), Pairwise distance features (PWD) combined with overlap index classifiers (OIC) and Local binary patterns (LBP) combined with histogram intersection classifiers (HIC) and also support vector machine (SVM) classifiers.

The PWD countermeasure was developed to detect converted voices and is based on the contraction of the cluster of features which occurs as a consequence of the conversion process. As expected, the countermeasure is largely effective in detecting spoofing attacks with voice conversion. Encouragingly, even when tested against other forms of spoofing attack for which it is not optimised, e.g. synthesized speech, the PWD countermeasure still provides acceptable performance in some cases.

The texture analysis countermeasure shows particularly encouraging performance for the mobile telephony scenario, giving almost perfect detection performance for artificial signals and speech synthesis, and also strong performance for voice conversion. We observe that an appropriate combination LBP features with one-class and two-class classifiers (i.e. the last two countermeasures shown in Table 8.5) would solve the problem of spoofing for the attacks and the setup considered in this thesis. This is however subject of further research.

On the other hand, initial experiments on speech signals of considerably reduced duration have shown that the efficiency of LBP-based countermeasures

is dependent on the length of the input signal. This can be explained by observing that a small number or speech features are insufficient to generate representative histograms[5] in the same way that a small number of speech features are insufficient to generate a representative GMM model [179]. Nevertheless, this poses a limitation for the practical use of LBP-based countermeasures and is subject of further research.

Finally, we observe that the combination of two independent classifiers (ASV system and countermeasure) makes assessment somewhat troublesome. While recently the EPS framework [48] appears as the first approach to address the problem of spoofing and countermeasures evaluations, this framework necessarily involves systems integration (ASV and countermeasure) at the score level while this thesis only consider setups integrated at the decision level.

The TABULA RASA evaluation methodology adapts methodologies for traditional ASV system evaluations (e.g. DET profiles) but we acknowledge that it is not an standardized approach with possible limitations/weaknesses and with the added difficulty to interpret the results. We nevertheless utilizes the TABULA RASA evaluation methodology simply because there is currently no alternative.

[5] we remain the reader that the LBP feature is basically formed by a concatenation of histograms

Chapter 9

Conclusions and Future Perspectives

The state-of-the-art in text-independent automatic speaker verification (ASV) has advanced rapidly in recent years. Surprisingly, however, there has been relatively little work in the development of countermeasures to protect such ASV systems from the acknowledged threat of spoofing.

This PhD thesis helps in the development of a new generation of more secure ASV systems able to provide reliable recognition. The thesis analyses ASV systems vulnerabilities and the mechanisms utilized by known spoofing attacks to fool them. It also searches for new threats and introduces new countermeasures that mitigate the effect of such threats.

More importantly, it acknowledges some fundamental weaknesses in the development of countermeasures, such as the improper use of prior knowledge, and establishes some of the foundations for further research, such as the use of Multiple classifier Systems (MCS) in future evaluation together with generalized countermeasures.

Due to the novelty of this research field there is still a long road ahead. Undoubtedly, the main issue relates to the lack of standard large-scale datasets, protocols or metrics which might otherwise be used to conduct evaluations on spoofing and countermeasures in a fairer sense and more comparable results, standards that are needed to be defined in the future. Further conclusions and future research directions are described in the following.

9.1 Vulnerabilities of ASV systems

In this thesis we are interested in the analysis and detection of *successful* spoofing attacks. In the context of this thesis, if an input signal not belonging to the targeted speaker (client) overcomes the speaker detection module **and** the speaker recognition classifiers of a given ASV system, then the signal is a

successful spoofing attack (Section 4.1).

The previous reasoning highlights the importance to access ASV vulnerabilities in this research. In this sense, from the vulnerability analysis presented in this thesis we observe that all the modules which comprises an ASV system present weak points that can be exploited by a spoofer.

As an example, Section 4.1 shows that energy-based SAD might be overcome by high-energy non-speech signals, that all feature representations can be mimicked and also speaker recognition classifiers can produce unpredictable outputs against non-speech signals for which they are not generally designed to cope with. For the former, these observations can thus be used as a warning for caution when using energy-based SAD, which might be preferred in practice over model-based of phoneme-based detectors (see Sections 2.2.1). However, there is no evidence that more advanced SADs are inherently more robust to spoofing.

The pursuing of feature representations that jointly improves recognition performance over the state-of-the-art i.e. MFCC and also enhances robustness against spoofing appears as a challenging task. On the other hand, weaknesses at the modelling stage are easier to identify, making the the latter a more likely direction of research. In this sense, this thesis present evidence which suggest that advanced algorithms, such as joint factor analysis or the i-vector scheme, may offer some inherent protection from spoofing.

This behaviour, contrary to our first hypothesis that channel compensation approaches may be of assistance to a would-be spoofer, is still unexplained and thus this line of research deserves more attention.

Score normalisation plays an ambiguous role in the face of spoofing. The results reported in Section 5.3.2 show significant differences in the FARs of the ASV systems with and without score normalisation when tested against the same attacks, but if score normalisation plays in favour or against spoofing depends of the attacks and the ASV considered.

However, we note that score normalisation may provide robustness against specific attacks (score normalisation provide consistently lower rates for white-noise) and specific systems (idem for IV-PLDA ASV system), being the latter an interesting observation since score normalisation is disregarded for i-vector schemes. In any case, further research is needed in this sense.

9.2 Spoofing attacks

The literature on spoofing is limited to four different attacks, including classical spoofing attacks such as impersonation or replay as well as more advanced attacks such as voice conversion and speech synthesis, all shown to provoke significant increases in the false acceptance rate of state-of-the-art ASV systems.

This thesis defines successful spoofing attacks by mean of a sufficient condition. The threats considered by this broad-sense, conservative definition of spoofing include but is not limited to the attacks mentioned above. To validate this approach this thesis addresses spoofing with non-speech, artificial signals.

Having the potential to pass both energy-based and pitch-based voice activity detection systems, artificial signals thus pose a serious threat to the reliability of ASV systems. For the tested systems and protocols, while voice conversion and voice synthesis attacks are not always as effective as artificial signals (depending on the operating point and ASV system used for artificial signal training) the threat is nonetheless significant. While artificial signals produce non-speech-like signals, voice conversion and voice synthesis have the advantage of producing speech-like signals. In line with previous, related work, this contribution highlights the importance of efforts to develop dedicated countermeasures, some of them trivial, to protect ASV systems from spoofing.

Results presented above show that all tested systems are vulnerable to spoofing through proposed attacks which provoke significant degradations in performance. We note, however, that results reported above are strictly related to specific techniques, systems and protocols which represent some, but certainly not all ASV system vulnerabilities. A full study of each spoofing attack represents a major research project in itself, i.e. many approaches to voice conversion and speech synthesis are reported in the literature and there are undoubtedly many more approaches to generate spoofing-specific artificial signals which we cannot address in the scope of this thesis. Therefore, while this work demonstrates the vulnerability of state-of-the-art systems it is only the first step towards a full understanding of the spoofing threat.

Finally, this thesis also addresses spoofing in terms of accessibility (effort). Except for replay attacks, all the addressed attacks are hardly reachable for the mass, either due of the need of specific skills (impersonation) or due of the need of specific expertise (voice conversion, speech synthesis and artificial signals). Further research should be focused in determine whether or not

ASV systems can be spoofed with other easily generated signals. These such attacks may be more representative of the practical spoofing scenario.

9.3 Countermeasures & integration

Numerous vulnerability studies suggest an urgent need to address spoofing and the solution seems not to be trivial. For instance the first countermeasure proposed for this thesis we hypothesized that a straightforward speech quality assessment routine, for example, may be used to distinguish artificial signals from genuine speech signals. The work reported in [12] shows the first study that utilized ITU-T specifications as speech quality assessment countermeasure, in this case against attacks with artificial signals. Against intuition, results are not satisfactory, which highlights the need of dedicated effort for countermeasure development.

This thesis reports three novel countermeasures for the protection of ASV systems from spoofing. The first two are a trivial approach based on repetitive pattern detection (RPD) and an approach based on pairwise distances analysis (PWD) to detect spoofing attacks with specific approaches to artificial signals and voice conversion, respectively. While in general the six tested ASV systems show considerable vulnerabilities against these two spoofing attacks, these two countermeasures are shown to be consistent and extremely effective in detecting spoofed attacks for which they were optimized.

To provide the countermeasures with some flexibility with respect to similar attacks (i.e. same attacks for which the countermeasures were designed but with generated with different spoofing configurations that the used in the laboratory), each countermeasure includes a one-class classifier trained only with real speech samples. Consequently, when tested against other forms of spoofing attack for which they are not optimized, both countermeasures still provide good performance in some cases. Yet, the overall result is not satisfactory.

While each of these countermeasures is successful in overcoming the specific attack considered, in reality system designers and countermeasure developers cannot assume such prior knowledge. In practice the spoofing attack can never be known and then the performance of existing countermeasures in practical scenarios cannot be guaranteed.

Accordingly, there is a need for generalized spoofing countermeasures with the potential to detect attacks for which they have not been optimized. This

thesis addresses this issue to some extent. The third proposed countermeasure is based on features obtained after local binary pattern (LBP) analysis of sequences of acoustic vectors. This feature, when combined with one-class classifiers, results in what is, to the best of our knowledge, the first generalised approach to spoofing detection for ASV systems.

Results show that the LBP-based countermeasure is less effective than specific solutions for voice conversion based spoofing attacks but that almost perfect detection is achieved for previously unseen spoofing attacks which otherwise provoke significant increases in false acceptance. Being less reliant on prior knowledge, the work points to the potential for generalized countermeasures with greater practical value.

Results also suggest that future work should consider the combination of specific and generalized countermeasures together with the recognition system/s they aim to protect (the latter due to the fact that in practice a countermeasure will never be used stand-alone). Combination of biometric systems and countermeasures is thus a key point in future research.

Combination of classifiers, known as Multiple Classifier Systems (MCS) theory, is also addressed in this thesis. Although in this thesis we do not carry out any research on fusion techniques to combine ASV systems and countermeasures, in Section 6.2 we introduce some of the basis to design MCS in the context of spoofing.

Part of the contribution of this thesis is related to the description of the different approaches to formulate the problem of reliable speaker verification (see Section 6.1). As fundamental as this task seems, it still remains without consensus among the researches in the biometric community. Possible approaches to problem formulation includes examples from the holistic approach which addresses spoofing to the conventional two-class problem to the reductionist approach which assigns one class per attack.

In particular, this thesis addresses spoofing and countermeasures formulated as a multi-class problem with outliers detection (Section 6.1.3.3). Evaluations in this thesis are reported by combining a maximum of two classifiers. Under this perspective, the combination of an ASV system with a generalised countermeasure, which gives the best overall results in this thesis, can be viewed as a 2-class problem with outliers detection.

Addressing spoofing and countermeasures as a multi-class problem with outliers detection appears as the most suitable approach given the state-of-the-art in the field. This approach addresses the appropriate use prior knowledge as

well as unseen attacks and generalized countermeasures (i.e. outliers detection) and integration of independent countermeasures to the recognition systems. However, to motivate research in this direction there is a clear need for formal spoofing and countermeasure evaluations. Formal evaluations with standard corpora, protocols and metrics are therefore needed to stimulate the research of spoofing countermeasures under properly controlled settings reflective of practical use case scenarios and with genuinely unseen and varying attacks. This discussion is addressed in the next section.

9.4 Evaluations & databases

Even if they stem from the adaptation of standard databases, all of the past work has been performed on non-standard databases of spoofed speech signals. This has usually entailed the development of a single, or small number of specific spoofing algorithms in order to generated spoofed trials. Countermeasure assessments are therefore biased towards the specific attacks and lack generality to new spoofing algorithms or entirely new forms of attack which will likely emerge in the future. Figure 6.3 shows that suitable evaluations play a critical role in the design cycle of MCS. In this sense, we note that current databases and evaluations are not representative of practical scenarios.

In order to address the inappropriate use of prior knowledge in future work, it will be necessary to collect and make available standard databases of both genuine speech and spoofed speech. Both the form of spoofing and the algorithms used to generates spoofed trials should include as much variation as possible in order to avoid bias and over-fitting. Standard databases will then encourage the integration of outliers detection (i.e. generalized countermeasures) to the ensemble of systems capable of detecting different, varying and perhaps previously unknown spoofing attacks which facilitate the meaningful comparison of different anti-spoofing countermeasures.

The design of spoofing datasets by adapting standard databases, although if preferred over small-scale studies involving purpose collected databases, may also not be rejective of some practical use-cases. As discussed in Section 4.3.2, with this setup attacks are simulated through post-sensor (or transmission) spoofing. This setup can be acceptable in the case of telephony applications, or if the sensor, channel and spoofing attack are all linear transforms but, in reality, this is unlikely. The setup is also unrealistic in the case of physical access scenarios where the microphone is fixed; the SRE data, for example, contains varying microphone and channel effects.

The main reason why attacks at the transmission level are much more studied that the sensor level counterpart related to the relative ease to generate the spoofing datasets. For instance, to simulate attacks at the sensor level, the spoofing utterances can be either re-recorded following specific protocols or either the spoofing utterances can be artificially convoluted with different impulse responses.

The latter approach appears far less troublesome to implement, although its validity should be investigated[1]. Further work should also investigate the relation between evaluations at the transmission level and sensor level (e.g. can evaluations performed at the transmission level help to infer results for evaluation at the sensor level?), insight in this field would enormously facilitate the design of spoofing datasets.

Furthermore, the majority of past work was also conducted under matched conditions, i.e. the data used to learn target models and that used to effect spoofing were collected in the same or similar acoustic environment and over the same or similar channel, whereas this might not be realistic. In order to reduce the bias in results generated according to such setups, future work should study the practical impact of the differences between the two experimental setups illustrated in Figure 4.4. Alternatively and preferably, future work should include the collection of new databases which more faithfully represent practical scenarios.

Finally, we observe that the combination of independent classifiers (e.g. ASV system and countermeasure) makes assessment somewhat troublesome. While recently the EPS framework [48] appears as the first approach to address the problem of spoofing and countermeasures evaluations, this framework necessarily involves systems combination (fusion) at the score level while this thesis only consider setups integrated at the decision level. The TABULA RASA evaluation methodology adapts methodologies for traditional ASV system evaluations (e.g. DET profiles) but we acknowledge that it is not an standardized approach with possible limitations/weaknesses and with the added difficulty to interpret the results. There is thus a need to develop standardised metrics to evaluate integrated countermeasure.

.

[1]Our work in [8] reports results with replay attacks in which the spoofing samples are artificially generated by the convolution of utterances with different impulse responses

9.5 Final thoughts

Spoofing and countermeasures are far from being a mature research field. The mechanisms behind impersonation are not fully understood and the constant improvements in high quality portable recorders make the detection of replayed speech a more challenging task (we note that there is almost no literature on countermeasures for these attacks). This thesis also suggests the potential to fool ASV systems with non-speech signals. Moreover, while voice conversion and speech synthesis technologies are in constant evolution, the current insight on spoofing leads to countermeasures to rely in excess on prior knowledge of the attack. Consequently, countermeasures are ineffective against unseen attacks.

Spoofing thus remains very much an open problem which appears not to have a definitive solution. More specifically, the problem of spoofing will last as long as the *effort* needed to overcome a ASV system encourage the would-be spoofer to do so (i.e. the effort needed to fool the system is less costly that the good the system aim to protect). This thesis stresses the assessment of spoofing in terms of effort and prioritize the study of low level spoofing attempts in future research.

A more interesting question relates to the **future direction** of this research. Independently of the biometric modality, this thesis infers that the next generation of biometric recognition systems will be implemented by mean of MCS which reflects the problem of reliable recognition formulated as a multi-class problem with outliers detection. For voice, due to the current insight on spoofing, the first ensemble of systems will probably be complex (i.e. by defining one class per attack) and with possible security breaches, but with the advance on the research in the field their complexity will decrease and gradually converge to the conventional two-class problem, in the hypothetical case where spoofing, able to be statistically modelled, is considered an intra-class variation. This question, as other several fundamental ones, is open to research, but taking into account that spoofing comprehends attacks going from impersonation to non-speech, artificial audio signals, this task is, at least, challenging.

Part III

APPENDICES

Appendix A

Evasion & Obfuscation in ASR systems

There is very little work in the literature relating to obfuscation, despite convincing arguments supporting the potential for obfuscation to overcome reliable recognition.

This appendix is an adaptation of the author's own work previously presented in [9, 10]. The work in [9] reports the first results on obfuscation over large-scale, standard databases (NIST) and is extended in [10] which reports the first investigation of evasion and obfuscation in the context of speaker recognition surveillance and forensics.

In contrast to spoofing, which aims to provoke false acceptances in authentication applications, evasion and obfuscation target detection and recognition modules in order to provoke missed detections. This appendix presents our analysis of each vulnerability and the potential for countermeasures using standard NIST datasets and protocols and six different speaker recognition systems (Section 5.1.1).

Results show that all systems are vulnerable to both evasion and obfuscation attacks and that the LBP-based countermeasure presented in Section 7.4 shows promising detection performance. While all evasion attacks and almost all obfuscation attacks are detected in the case of this particular setup, the work nonetheless highlights the need for further research.

A.1 Introduction

While spoofing research in automatic speaker verification (ASV) is only just beginning to gather pace [66], there is almost no work in the literature related to either evasion or obfuscation [97]. Reliable recognition performance is essential whatever the application (Section 1.1). It is thus essential that studies of evasion and obfuscation are made in parallel with, and to accelerate

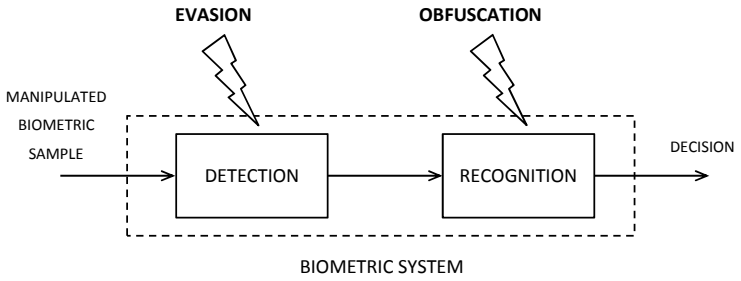


Figure A.1: An illustration adapted from Figure 4.1 to show scenarios for evasion and obfuscation in biometric recognition.

the design of new approaches to detect manipulated evidence and to ensure the reliability of automatic speaker recognition in surveillance and forensic applications.

This appendix reports our study of evasion and obfuscation in the context of ASV. The potential to provoke missed detections is assessed using six different ASV systems, from a standard GMM-UBM system to a state-of-the art i-vector system. New to this contribution is the classification and study of independent evasion and obfuscation attacks. Also this appendix reports the results obtained when using the LBP-based countermeasure to identify attempts to evade and obfuscate detection.

A.2 Evasion and obfuscation

Both evasion and obfuscation refer to the intentional manipulation of a biometric signal in order to provoke missed detections. While the notion of obfuscation is now widely understood [207] we see fundamental differences between evasion and obfuscation; while the end result is the same, the attacks target two distinctly different components of a typical biometric system.

Evasion and obfuscation attacks are illustrated in Figure A.1. It shows the two critical elements in a standard biometric system, namely the detection and recognition modules, which may be vulnerable to evasion and obfuscation respectively.

A.2.1 Evasion

The detection module aims to identify those components or intervals of the input signal which are of interest to the recognition module, i.e. typically the components containing a face, a fingerprint, or the intervals containing speech. Evasion attacks can be applied here to prevent such components from being identified. Consequently, the recognition module will never receive a valid biometric sample.

In terms of ASV, the biometric detector is commonly referred to as either speech activity detection (SAD) or voice activity detection (VAD) (see Section 2.2.1). While model-based and phoneme-based approaches are more complex and conceivably more robust to evasion, energy-based approaches can be overcome with relative ease. Since they generally rely on relatively clean signal-to-noise ratios, a simple attack might involve the filling of non-speech periods with a signal whose energy is higher than that of the speech. As a result, only non-informative intervals which do not contain speech will then be passed to the recognition module.

Energy-based SADs are still well accepted in the literature and, mostly for reasons of computational simplicity, they might be preferred in practice. There is also some recent evidence [172] which suggests that energy-based SAD can be more effective than alternative model and phoneme-based SAD in a variety of noise conditions.

A.2.2 Obfuscation

Assuming that useful speech does reach the recognition stage, then here there is potential for the speech signal to be manipulated in order to interfere with the decision and once again provoke a recognition error. In line with the definition for fingerprint recognition [207] we refer to speech obfuscation as the intentional manipulation of an utterance in order to provoke missed detections.

In this context, obfuscation can be seen as a sub-domain of voice disguise, which considers both intentional and non-intentional speech alterations [164]. Other approaches might include automatic manipulations such as voice transformation and voice conversion [154], pitch modification (e.g. falsetto), whispering, glottal fry, pinched nostril speech, bite blocking, a hand over the mouth, imitation and other mechanical/prosody alterations.

There is very little work in the literature relating to obfuscation, despite con-

vincing arguments supporting the potential. The work in [113, 100, 211, 192] investigated the effect of intentional voice modifications or disguise and found in all cases that missed detection rates increase. Automatic approaches to voice transformation reported in [97, 155] are also shown to overcome identification and verification systems, though most of this work involves the use of non-standard, small datasets. The first work to detect disguised voice is reported in [196]. While performed using the standard TIMIT database and while promising detection rates are reported, the work does not consider impacts on ASV performance.

This appendix presents the first assessment of evasion and obfuscation under controlled conditions using large-scale, standard NIST databases and state-of-the-art approaches to obfuscation which have already been shown to overcome ASV through spoofing. We also present a new approach to evasion and obfuscation detection and analyse its impact on ASV performance.

A.3 Evaluation

This section presents our work to assess the vulnerability of automatic speaker verification (ASV) to evasion and obfuscation. We describe the different ASV systems, datasets and protocols used in this work, the particular approaches to evasion and obfuscation, and experimental results.

We stress that the full consideration of every possible threat is beyond the scope of this contribution. Clearly this work is only a start to broader research which will require greater attention in the future.

A.3.1 Experimental setup

We assessed the impact of evasion and obfuscation on six different ASV systems: (i) a standard GMM-UBM system; (ii) a GMM-UBM system with factor analysis (FA) channel compensation; (iii-v) three different GMM supervector linear kernel (GSL) systems, and (vi) a state-of-the-art i-vector system. Details of system configurations can be found in Section 5.1.

All development was performed using the male subset of the 2005 NIST Speaker Recognition Evaluation dataset (NIST‘05) whereas the male subset of the NIST‘06 dataset was used for evaluation. Only evaluation results are reported in this appendix. The NIST‘04 or NIST‘08 datasets are used as

background data, depending on whether the data is used for ASV or evasion and obfuscation respectively.

To assess the potential impact of evasion and obfuscation, true-client tests are replaced with alternative speech data which aims to either evade or obfuscate reliable recognition. Any number of different approaches may be used. The specific approaches chosen in each case are described in the next sections. The only difference between their use in the study of evasion and obfuscation instead of spoofing involve their application to client trials (instead of impostor trials) to provoke missed detections (instead of false accepts).

The scale of the evasion threat will naturally depend on the specific approach to SAD. In the following, we assume the use of a simple, energy-based approach and report illustrative examples with a equally straightforward, targeted attack in order to demonstrate the concept.

The input speech signal is first processed offline to identify low-energy, neighbouring intervals of non-speech. The average energy level of the intervals containing speech is then estimated and the non-speech intervals alone are filled with higher-energy white noise. While the resulting signal is perceptually challenging, and with the exception of some masking effects, the speech remains entirely intelligible.

As described in Section A.2.1, only the higher-energy components of the input signal are retained after SAD. Such a trivial attack thus succeeds in ensuring that very little, if any useful clean speech is passed to the speaker recognition system which instead receives only intervals of non-informative noise.

The approach to obfuscation used in this work is based on voice conversion. It is applied here according to the Gaussian dependent filtering approach proposed in [133] and described in Section 3.2.4.2.

In order to simulate obfuscation, voice conversion is applied to all true-client test utterances. To increase the chances of provoking missed detections we further convert each utterance towards the most dissimilar speaker among a selection of 10 randomly chosen subjects (that for which the likelihood score from a conventional trial is the lowest).

A.3.2 Results

Baseline ASV results are presented together with those for evasion and obfuscation in Table A.1. Results are illustrated in terms of the equal error rate

(a) ASV systems without score normalisation.

System/Attack	ASV			ASV + CM	
	-	Evas	Obf	Evas	Obf
GMM-UBM	8.7	19.4	47.7	0	4.3
SGL	8.0	55.1	32.3	0	3.5
SGL-NAP	6.8	53.4	31.5	0	3.3
SGL-FA	6.4	54.7	29.1	0	3.3
FA	5.6	20.6	41.9	0	3.5
IV-PLDA	3.0	24.3	20.0	0	3.8

(b) ASV systems with score normalisation.

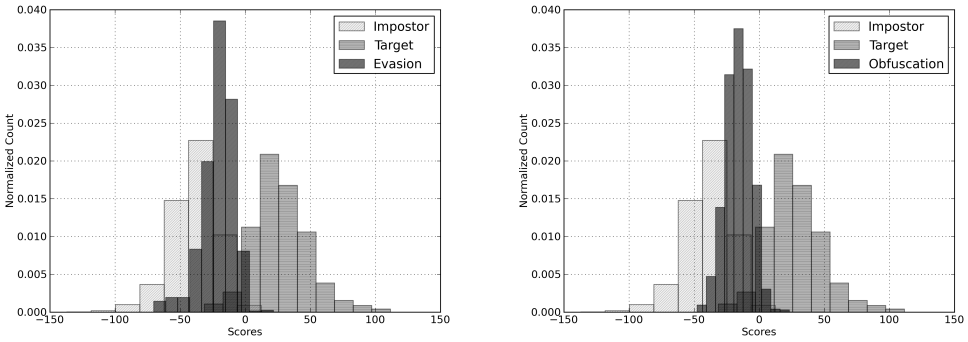
System/Attack	ASV			ASV + CM	
	-	Evas	Obf	Evas	Obf
GMM-UBM	8.6	52.6	32.1	0	4.5
SGL	8.1	53.2	29.9	0	3.5
SGL-NAP	6.3	50.3	27.6	0	3.5
SGL-FA	5.7	49.8	32.4	0	3.4
FA	5.6	49.1	29.2	0	3.8
IV-PLDA	2.9	49.8	26.8	0	3.3

Table A.1: ASV performance without speech alteration (baseline), with evasion with noise and obfuscation through voice conversion. This analysis is repeated from the fourth to the sixth columns for the ASV system with integrated LBP countermeasure. Results shown in terms of EER (%) and for ASV systems with and without score normalization.

(EER), score distributions and DET profiles (the latter two only for PLDA systems). We discuss only the former in the following.

Table A.1(a) shows the ASV system without normalised scores under evasion. For GSL-based systems, the EER increases from in the order of 7% to over 50%. On the other hand, the baseline EERs for the GMM-UBM, FA and IV-PLDA systems increase from between 3% and 9% to between 19% and 24%. When the scores of the six studied ASV systems are normalised (Table A.1(b)) the EER is in the order of 50% in all cases.

Table A.1 also illustrates the effect of obfuscation. Results show that for ASV system without score normalisation the GMM-UBM system is the most vulnerable; the EER increases from 9% to 48%. The FA and three GSL-based systems also show high levels of vulnerability (EERs between 29% and 42%), whereas the IV-PLDA system is the most robust; the EER increases from 3%



(a) Evasion (b) Obfuscation

Figure A.2: IV-PLDA score distributions for impostor (left-most) and target (right-most) are common to both Figure A.2(a) and Figure A.2(b). Figure A.2(a) shows evasion trials with white noise while Figure A.2(a) shows obfuscation trials with voice conversion.

to 20%. When the scores of the six studied ASV systems are normalised the EER is in the order of 30% in all cases.

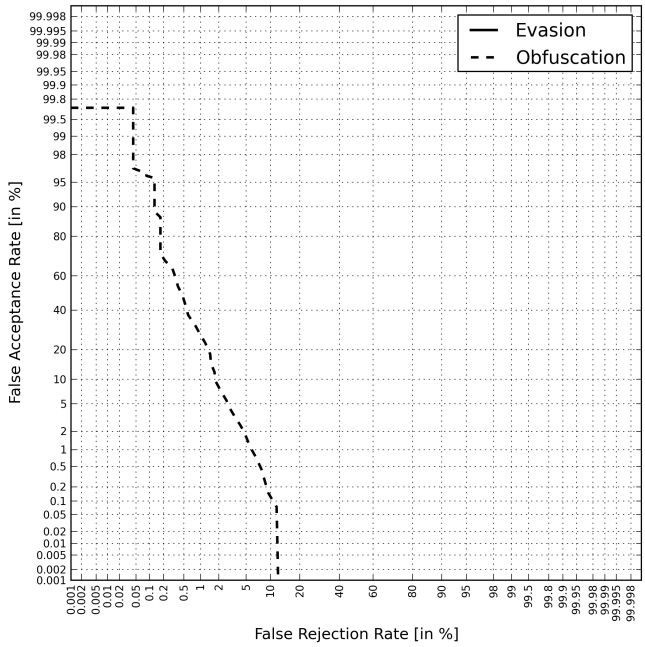
Figure A.2 shows a histogram of scores for impostor trials (left-most distribution) and target trials (right-most) for the IV-PLDA ASV system. Also illustrated is the score distribution for evasion (Figure A.2(a)) and obfuscation tests (Figure A.2(b)) which shows how white noise and voice conversion respectively are effective in decreasing the likelihood scores for target tests; the degree of overlap with the impostor distribution is higher than for the target distribution thus accounting for the increase in EER.

Detection error trade-off (DET) profiles for the IV-PLDA system are illustrated in Figure A.3(b). Profiles for the baseline and obfuscation show that the system is vulnerable across the full range of operating points.

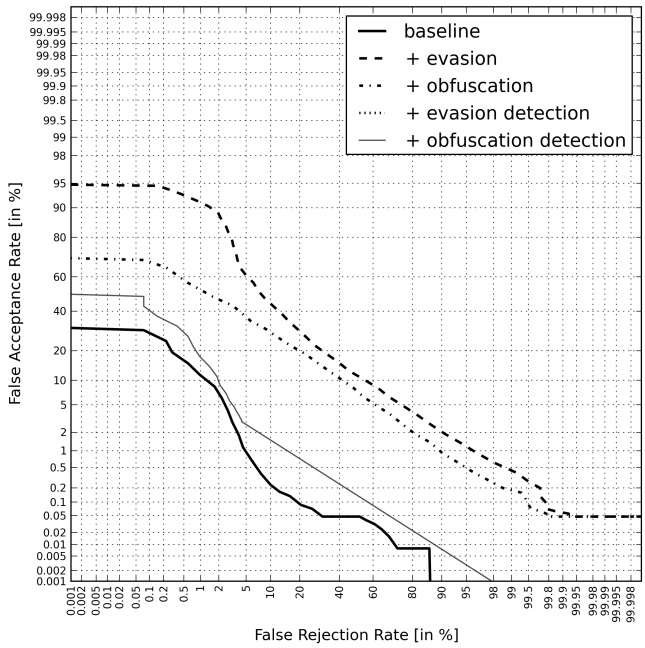
A.4 Detection

Various different approaches to detect manipulated speech signals have been reported in the literature. All involve the study of spoofing and the detection of processing artifacts indicative of manipulation, e.g. the absence of natural-speech phase [198] and reduced short-term dynamic variability [6].

These approaches are, however, dependent on the specific approach to spoofing and thus have limited practical application. The work in [5] presented the first



(a) LBP-based countermeasure evaluated in stand-alone operation



(b) LBP-based countermeasure integrated with IV-PLDA ASv system

Figure A.3: DET profiles illustrating the performance of the LPB-based countermeasure evaluated independently of the ASV system (Figure A.3(a)) and the performance of the IV-PLDA system for the baseline, the baseline with obfuscation and evasion attacks and then with integrated detection (Figure A.3(b)).

generalised solution with the potential to detect previously unseen approaches to manipulation.

A new, one-class classification approach learnt using only genuine speech is used to detect the absence of natural spectro-temporal variability through the so-called local binary pattern (LBP) analysis of speech spectrograms. With improved generalisation, this approach to detection has greater practical application and is thus the approach adopted here as a means of detecting both evasion and obfuscation.

DET plots illustrating detection performance in independence from ASV for both evasion and obfuscation are illustrated in Figure A.3. The EER for evasion detection is 0% whereas that for obfuscation detection is 3%. ASV performance with combined obfuscation detection as a post-processing step [6] is illustrated for the IV-PLDA system in Figure A.3(b). With the detector operating point set to the EER (Figure A.3(a)) there is almost no degradation in ASV performance (Figure A.3(b)) towards the low missed detection region. Corresponding EERs with integrated detection for all six systems are also illustrated in Table 1 and show EERs in the range of 3 to 4% in all cases.

A.5 Conclusions

This appendix demonstrates the potential for surveillance and forensic speaker recognition systems to be manipulated. While ultimately they have the same effect, we introduce the notion of different, independent vulnerabilities to evasion and obfuscation which target either the detection or recognition modules. More importantly, this work demonstrates the need and potential for new evasion and obfuscation detection countermeasures.

We acknowledge that the work presented in this appendix is far from being exhaustive. Even if the trivial form of evasion examined in this work may not overcome more sophisticated speech activity detection systems, and while the approach to obfuscation is perhaps beyond the means of the lay person, the observations reported here serve to highlight the need for further research to ensure that surveillance and forensic systems are adequately protected from both forms of subversion.

Appendix B

Résumé Etendu en Français

Cette section présente un résumé de la thèse écrite en français. En raison de l'impossibilité de faire une synthèse équitable de tous les contenus de la thèse, l'auteur de cette thèse a choisi une méthodologie qui est présenté comme suit.

La version française de cette thèse comprend les contributions à la recherche (Section B.1.2) et conclusions (Section B.4) sans modifications de la version anglaise. Sections Section B.2 et Section B.3 présentent une synthèse des principaux résultats des évaluations de spoofing et les contre, respectivement. Les informations manquantes à partir de la version anglaise est reconnaître lorsque ce est nécessaire.

B.1 Introduction

La biométrie fait référence aux technologies qui mesurent et analysent les caractéristiques physiologiques et/ou comportementales d'une personne comme les empreintes digitales, de l'iris, de la voix, la signature manuscrite, le visage, l'ADN, la démarche, la géométrie de la main et d'autres pour la reconnaissance de leur identité (vérification ou identification). La biométrie offre une alternative par rapport aux méthodes traditionnelles de reconnaissance de personne, en s'appuyant sur *ce que vous êtes* ou *ce que vous faites* par opposition à *ce que vous savez*, comme un numéro de code PIN ou mot de passe, ou *ce que vous avez* comme une carte d'identité, un jeton ou un passeport.

Á partir du nombre de différentes biométries, la reconnaissance de l'identité d'une personne utilisant leur voix suscitent un intérêt comme aussi importante; les signaux de la voix sont facilement capturés dans presque n'importe quel environnement en utilisant des microphones standard et du matériel d'enregistrement, y compris à distance, comme par exemple, par téléphone, où la parole est souvent le seul mode biométrique qui est disponible. Depuis, leurs attraits naturel réside dans des scénarios automatisés et sans surveillance, les systèmes de reconnaissance du locuteur sont particulièrement vulnérables à le leurrage (attaques de type spoofing).

La vérification automatique du locuteur (VAL) est un domaine de recherche mature. Cependant, en comparaison avec d'autres modalités biométriques, le leurrage (désormais notée comme spoofing) et la recherche de contre-mesure dans les systèmes VAL sont beaucoup moins avancés.

Cette thèse présente quelques-unes des premières solutions à ce problème. En particulier, il répond à certaines des questions découlant du spoofing afin de faire confiance à des systèmes VAL.

B.1.1 Motivation des contre-mesures contre le spoofing

Aussi dénommé comme attaques directes ou attaques aux niveau du capteur de systèmes biométriques [69], le *spoofing* se réfère à la présentation d'un trait falsifié ou manipulé au capteur d'un système biométrique pour provoquer la validation illégitime. À moins que le système biométrique ne soit équipé avec les contre-mesures appropriées, cette menace est commune à toutes les modalités biométriques. Par exemple, les systèmes de reconnaissance de visage peuvent être falsifiés avec une photo [63], tandis que les systèmes de reconnaissance d'empreintes digitales ou vocales peuvent être falsifiés avec un faux doigt en gomme [79] ou avec un enregistrement audio [192], respectivement.

Les systèmes de sécurité doivent être constamment mis à jour. Un système qui est supposé être sûr de nos jours peut devenir obsolète si il n'est pas régulièrement améliorée. C'est particulièrement vrai pour la biométrie, pour laquelle la garantie de fiabilité est une exigence cruciale pour l'adoption continue des systèmes biométriques dans le marché de la sécurité.

Cette thèse apporte quelques idées ou problème de la reconnaissance fiable pour les systèmes VAL par l'analyse et l'évaluation de leurs vulnérabilités contre le spoofing, l'enquête de courant (et nouvelles)menaces ainsi que le développement de nouvelles contre-mesures qui atténuent les effets de ces menaces.

B.1.2 Contributions

La liste suivante montre les contributions de recherche apparaissant dans cette thèse. La liste fait référence également au travail réalisé comprenant neuf articles publiés lors de conférences (**C1**-**C9**), trois chapitres de livres (**B1**-**B3**) et un journal (**J1**) résumés dans Liste des Publications.

- Nouvelles attaques de type spoofing

  - Une nouvelle approche pour les signaux artificiels est d'abord signalée dans (**C1**) et aussi dans la Section B.2.2.2. Au mieux de notre connaissance, ce travail est le premier à considérer la vulnérabilité potentielle des systèmes VAL aux signaux non vocaux. En outre, les résultats expérimentaux mentionnés dans la Section B.2.4.2 montrent que les attaques avec des signaux artificiels sont une menace pour les systèmes VAL.

  - Les résultats expérimentaux présentés dans la Section B.2.4.2 suggèrent que les attaques avec des signaux de bruit blanc (attaques de faible effort) sont sans doute une menace plus grave par rapport à des imposteurs naïfs.

- Nouvelles idées sur les vulnérabilités du système VAL et attaques de type spoofing

  - Nos expériences avec le bruit blanc, ensemble, avec quelques observations dans la littérature (c.-à-d l'existence des soi-disant *loups* [35, 61] -locuteurs imposteurs qui ont un potentiel naturel d'être confondus avec d'autres) conduit à la notion d'attaques de type spoofing généralisées (bien que l'existence généralisée des attaques par spoofing n'st pas évaluée dans cette thèse).

  - La Section B.2 rapporte la première étude sur la vulnérabilité contre le spoofing avec une attention particulière dans l'effet de la normalisation de score. La relation entre la normalisation de score et le spoofing est également abordée dans la Section B.2.5

- Nouvelles contre-mesures

  - Trois nouvelles contre-mesures sont résumées dans cette section. La première a trait à la détection de signaux artificiels et est présenté dans (**C2**), la deuxième est une approche spécifique pour détecter les voix convertis et est présenté dans (**C3**), la troisième est une approche généralisée avec deux variantes, présentées dans (**C4**) et (**C5**), respectivement.

  - Cette thèse présente la première approche généralisée de détection des attaques de type spoofing parmi toutes les modalités biométriques. Cette approche, basée sur la classification "1-classe", est d'abord présenté dans (**C5**) puis aussi dans (**J1**) et la Section B.3.1.3.

- Première évaluation des contre-mesures dans un framework multi-système, multi-attaque

  - Les études antérieures sur les évaluations de la vulnérabilité contre les attaques de type spoofing sont pour la plupart dans des conditions spécifiques, à savoir, un scénario, une attaque et un système VAL. Cette thèse présente la première étude comparative avec une large gamme de systèmes VAL y compris le système état de l'art i-vecteur avec post-traitement PLDA.

  - Notre travail dans (**C4**) est la première qui évalue les contre-mesures dans un environnement multi-attaque et multi-systèmes commune, pour les attaques de nature différente, y compris la conversion de la voix, la synthèse de la parole et des signaux artificiels, respectivement. La partie expérimentale de cette thèse évaluer les contre-mesures pour six systèmes VAL différents, y compris le système état de l'art i-vecteur avec post-traitement PLDA et les trois attaques mentionnées.

Contributions élaborées dans le cadre de cette thèse et inclus dans la version anglaise de ce document:

- Nouvelle relecture de la littérature sur les attaques de type spoofing et contre-mesures

  - L'auteur de cette thèse a participé a des relectures de la littérature inclus dans (**B1**), (**B2**), (**J1**), (**D1**) et (**D2**). Il les a adaptés et inclus dans la Partie I du présent document.

- Nouvel aperçu des vulnérabilités du système VAL et évaluations

  - Une analyse des vulnérabilités VAL qui complète le travail dans (**B1**) et (**J1**) est rapporté dans la Section 4.1

  - Ce travail réévalue le spoofing en termes d'effort. Il catégorise les menaces dans les attaques d'effort faible, moyen et haute ainsi que les attaques zero-effort connus (imposteurs naïfs).

  - L'auteur aborde le problème de l'évaluation de la vulnérabilité pour les modalité de la voix d'une discussion présentée dans la Section 4.3 et (**B3**)

- Nouvel aperçu sur le problème de la vérification biométrique fiable

  - Un cadre théorique sur le problème de la reconnaissance fiable du locuteur est présenté dans la Section 6.1. Au mieux de notre connaissance, ce travail est le premier à formuler des contre-mesures

et le spoofing comme un problème multi-classes avec détection des valeurs aberrantes (outliers).

- Le problème de l'évaluation et de l'intégration d'une contre-mesure est discutée dans la Section 6.2. En particulier, le problème de l'intégration de contre-mesure est traitée dans le contexte de "systèmes a multiple classeurs" (SMC, ou MCS en anglais).

- Première étude de l'évasion et de l'obscurcissement avec des bases de données standard a grande échelle
  - Cette thèse est la première à réévaluer le problème de l'obscurcissement par la classification et l'étude des attaques perpétrées indépendamment au niveau de détection de la biométrie ainsi qu'au niveau de la reconnaissance qui sont redéfini comme l'évasion et l'obscurcissement, respectivement. Ce travail est d'abord signalée dans (**C7**).
  - Les travaux rapportés dans (**C6**) et (**C7**) sont les premiers à rapporter les résultats sur l'obscurcissement avec des bases de données standard a grande échelle (NIST).
  - Les travaux rapportés dans (**C7**) montre que la contre-mesure à base de motif binaire local (MBL ou LBP en anglais) peu détecter l'évasion et l'obscurcissement avec une précision raisonnable.

Autres contributions élaborées dans le cadre de cette thèse, mais ne figurent pas dans ce document:

- Premier étude d'attaques "replay" avec des bases de données standard à grande échelle
  - Notre travail dans (**C8**) réévalue la menace d'attaques "replay". Les résultats montrent que, malgré le manque d'attention de ces attaques dans la littérature, des attaques "replay" de faible effort représentent un risque important, dépassant, comparativement, celles des attaques de haute effort telles que la conversion de la voix et la synthèse de la parole.

- Première étude comparative des attaques en scénarios l'accès physique
  - Base de données EURECOM: une base de données style MOBIO avec environ 18 heures d'échantillons audio/vidéo de 21 sujets (1260 échantillons provenant de 14 hommes et 7 femmes) ont été recueillis pour évaluer le spoofing dans les scénarios d'accès physiques. Les

détails et protocoles sont présentés dans (**D3**).

- Un système VAL basé en GMM-UBM est évaluée pour le scénario d'accès physique (base de données EURECOM) et pour quatre attaques différentes, y compris l'attaques "replay", la conversion de la voix, la synthèse de la parole et des signaux artificiels. Les résultats sont présentés dans les rapports (**D4**) à (**D7**).

- Nouveaux travaux sur les contre-mesures
  - Les travaux rapportés dans (**C2**) montrent la première étude ayant utilisé les spécifications de l'ITU-T (évaluation de la qualité de la parole) comme contre-mesure, dans ce cas, contre les attaques avec des signaux artificiels. Contre toute intuition, les résultats ne sont pas satisfaisants, ce qui ouvre la discussion sur la nécessité d'un effort soutenu pour le développement de contre-mesures.
  - Une contre-mesure en fonction des motifs binaires locaux (MBL) de une dimension est présenté dans (**D6**).
  - Le premier travail expérimental qui montre les résultats liés à la fusion de deux contre-mesures est rapporté dans (**D6**)

B.2 Évaluation des attaques de type spoofing

Cette section présente une synthèse de notre analyse pour mesurer l'efficacité des attaques directes à les systèmes VAL et de questions connexes, afin de fournir un aperçu de la vulnérabilité des différents systèmes de reconnaissance contre ces menaces. Cette section définit les spécifications relatives à la performance des systèmes VAL et l'évaluation de vulnérabilité, y compris les description et la configuration des systèmes VAL et attaques de type spoofing, bases de données biométriques et de spoofing, et les protocoles et les mesures adoptées pour chaque évaluation.

B.2.1 Spécifications pour l'évaluation de base

Dans ce qui suit, nous nous concentrons sur les résultats de base (baseline) liées à les bases de données des campagnes d'évaluation de NIST-SRE. Ils sont basés sur la téléphonie et concernent sans doute l'utilisation la plus attrayante de la reconnaissance vocale, à savoir la reconnaissance à distance par téléphone.

Le scénario est l'un des plus difficiles en termes de spoofing et contre-mesures, car il est entièrement non supervisé et est donc particulièrement vulnérable aux attaques de spoofing.

B.2.1.1 Base de données de base

Nous visons à évaluer l'effet des attaques de type spoofing sur une gamme de systèmes basés sur les développements récents dans le domaine de la reconnaissance automatique du locuteur. Toutes les conduire à la performance d'état de l'art à en juger par les campagnes d'évaluation de NIST[1] et sont tous basés sur la plateforme ALIZE [33]. L'inclusion de plusieurs systèmes de base dans le cas de reconnaissance du locuteur est motivée par l'impact probable de différents algorithmes de compensation de canal qui peuvent être utiles à un probable usurpateur (spoofer). Comme c'est une pratique courante, distincts systèmes sont indépendamment pour des genres masculin et féminin. Nous nous concentrons uniquement sur le bases de donnes de genre masculin dans cette thèse. En résumé, les bases de données NIST'04 et NIST'08 sont utilisés comme donnes auxiliaire (background), le base de données NIST'05 sont utilisés pour le développement et le base de données NIST'06 est utilisé pour l'évaluation. Toutes les expériences se rapportent à la condition de base (core) (1conv4w-1conv4w) qui implique environ 2,5 minutes de données pour la formation de modèle et les tests et tous les systèmes sont optimisés en fonction de taux d'erreurs égales EER (Equal Error Rate) avec des performances dynamiques évaluées selon les courbes DET (trade-off error). A noter que, dans le cas des bases de données NIST-SRE, c'est en contraste avec convention qui dicte l'optimisation en fonction de la métrique minDCF (minima de la fonction de coût de détection).

B.2.1.2 Configuration des systèmes VAL

Des expériences ont été menées avec six systèmes VAL. Ils sont tous basés sur l'outil LIA-SpkDet [31] et la bibliothèque ALIZE [32] et proviennent directement de le travail dans [72]. Dans tous les cas le signal de parole est divisé en fenêtres de 20 ms avec à décalage régulier de 10 ms. Dans tous les cas les systèmes utilisent un paramétrage commun où caractéristiques extraites en utilisant SPro sont composés de 16 coefficients cepstraux de fréquence linéaire (LFCC), leurs dérivées premières et le delta de l'énergie. Un système commune de détection d'activité vocale basée sur l'énergie (SAD) est également utilisé

[1] http://www.itl.nist.gov/iad/mig/tests/spk/

pour supprimer des fenêtres de non-parole et tous les systèmes utilisent un modèle commun universel de fond (UBM) avec 1024 gaussiennes. Le premier système VAL est un système standard modèle de mélange gaussien (GMM) avec un modèle du monde (UBM) notée GMM-UBM. Le deuxième système est un classificateur machine à vecteurs de support (SVM) qui est appliqué à supervecteurs GMM provenant directement du système GMM-UBM. Il est considéré comme un système GSL (supervector linear kernel system [39]). Le troisième système est presque identique à le deuxième, mais est améliorée grâce à la technologie NAP (nuisance attribute projection [40]) pour atténuer la variabilité intercanal, avec des matrices de NAP de rang 40. Le quatrième système comprend compensation de canal basée sur l'analyse des facteurs (FA) [101] selon l'approche symétrique présenté dans [135]. La cinquième approche est un système GSL avec supervecteurs FA (GSL-FA) [72] et le sixième est un système i-vecteur, le système état de l'art dans la vérification du locuteur [59], qui comprend des mélanges 1024 composants Gaussienes et des i-vecteurs de dimension 400. La variabilité indésirable est manipulé par la compensasion par le analyse discriminante linéaire probabiliste (PLDA) [118] avec d'une normalisation de longueur [81].

Pour les systèmes i-vecteur, en raison de la quantité importante de données nécessaires pour estimer la matrice de variabilité total T, le base de données NIST'06 a été utilisé au cours du développement et de le base de données NIST'05 a été utilisé pendant l'évaluation. Dans les deux cas, les bases de données complétées par les bases de données et NIST'04 et NIST'08 ont été également utilisé. Dans les deux cas, les matrices sont estimés à environ 11000 enregistrement de 900 locuteurs, tandis que l'indépendance entre les expériences de développement et d'évaluation est toujours respectée.

Tous les systèmes ont été testés avec et sans l'application de T-norme (à l'exception du système i-vecteur qui utilise S-norme) en utilisant des imposteurs de la base de données NIST'04.

B.2.2 Spécifications pour l'évaluation de vulnérabilité

Cette section décrit le framework expérimental utilisé pour évaluer le spoofing dans cette thèse. Pour la biométrie vocale nous allons enquêter sur les attaques avec conversion de voix, signaux artificiels et de synthèse de la parole, tandis que les attaques de relecture ne seront pas évaluées dans cette thèse.

Cette section décrit la configuration des attaques de type spoofing ainsi que la conception des bases de données de type spoofing et les protocoles pour les

transactions licites et l'évaluation de le spoofing.

B.2.2.1 Bases de données spoofing

Le travail expérimental sera abordée dans cette thèse en utilisant bases de données NIST-SRE en utilisant exactement les mêmes ensembles de données que celui utilisé pour les évaluations de base comme indiqué dans la section B.2.1. Plus de détails concernant les spécifications de chaque jeu de données est disponible gratuitement sur le site Web du NIST. Toutes les données utilisées pour effectuer des attaques par spoofing proviendront des mêmes ensembles de données.

La seule différence entre les données utilisées pour les études de base, rapporté à la section B.2.1 et les données à utiliser pour le spoofing se rapporte à l'utilisation d'une condition expérimentale différente. Au lieu de la condition de 1conv4w utilisés pour des études de base, le spoofing et les études connexes seront effectué sur la condition 8conv4w qui fournit plusieurs sessions pour chaque locuteur. Il contient des enregistrements vocaux de 201 et de 298 locuteurs pour NIST'05 et NIST'06 (masculin), respectivement. Il y'a 8 séances pour chaque locuteur pour un total de $499 \times 8 = 3992$ enregistrements d'une durée d'environ 2,5 minutes. Ces séances ne sont pas utilisés pour les tests sont utilisés pour effectuer des attaques par spoofing pour tous les essais imposteurs dans les protocoles standards NIST.

En plus de la nouvelle base de données de référence (sans spoofing), trois nouveaux ensembles de données seront générés où tous les segments de test imposteur sont remplacés par des versions falsifiés provenant de signaux artificiels, conversion de voix ou de synthèse vocale.

B.2.2.2 Description & configuration des attaques de type spoofing

La configuration des systèmes de spoofing sont réalisées selon deux hypothèses importantes. Se il est certes pas représentatif de scénarios réels, nous évaluons la performance de contre-mesure dans le pire des cas, où l'attaquant/usurpateur a connaissance préalable complète c.-à-d. la technologie de système de VAL utilisée, la configuration du système VAL, etc. D'autre part, nous conservons les données utilisées pour apprendre le système de spoofing (c.-à-d. données d'apprentissage et des données auxiliaires) indépendant des données utilisées dans les systèmes VAL de base.

Conversion de Voix

Tous les travaux impliquant la conversion de voix ont été réalisée avec notre propre mise en œuvre de l'approche proposée à l'origine dans [133]. Il a été développé pour tester les limites des systèmes VAL lorsque l'information de conduit vocal dans le signal de parole d'un imposteurs est convertie vers celui d'un autre. Au niveau du fenêtre, le signal de parole d'un imposteur notée $y(t)$ est filtré dans le domaine spectral de la manière suivante:

$$Y'(f) = \frac{|H_x(f)|}{|H_y(f)|} Y(f) \tag{B.1}$$

où $H_x(f)$ et $H_y(f)$ sont les fonctions de transfert du conduit vocal du locuteur ciblé et l'imposteur, respectivement. $Y(f)$ est le signal de parole de l'imposteur tandis $Y'(f)$ désigne le résultat après la conversion de voix. En tant que tel, $y(t)$ est mappé ou converties vers le locuteur cible dans un sens spectrale-pente. Comme nous le verrons plus loin, c'est suffisant pour surmonter la plupart des systèmes VAL.

$H_x(f)$ est déterminée à partir d'un ensemble de deux modèles de mélange gaussien (GMM). Le premier, noté comme de modèle de reconnaissance automatique du locuteur (asr) dans le travail originale, est liée à l'espace de caractéristiques VAL et utilisés pour le calcul des probabilités a posteriori, tandis que le deuxième, noté que le modèle de filtrage (FIL), est un modèle lié des coefficients LPCC dont $H_x(f)$ est dérivé. Paramètres de filtres LPCC sont obtenus selon:

$$x_{fil} = \sum_{i=1}^{M} p(g^i_{asr}|y_{asr}) \mu^i_{fil} \tag{B.2}$$

où $p(g^i_{asr}|y_{asr})$ est la probabilité a posteriori de la composante gaussienne g^i_{asr} étant donné y_{asr} et μ^i_{fil} est la moyenne du composant g^i_{fil} qui est liée à g^i_{asr}. $H_x(f)$ est estimée à partir de x_{fil} utilisant une transformation de LPCC à LPC et un signal temporel est synthétisé à partir d'images converties avec une technique *overlap-add*. Tous les détails peuvent être trouvés dans [133, 29, 30].

En raison de l'hypothèse de la pire des cas, le traitement *front-end* utilisé dans la conversion de voix est donc exactement le même que celui utilisé pour les systèmes VAL. Le modèle de filtrage g_{fil} et filtrer $H_x(f)$ utilise 19 LPCC et LPC (coefficients alpha), respectivement, calculée avec SPRO.

Signaux artificiels

Les attaques de signal artificiel sont basés sur l'algorithme indiqué dans [14]. Il est basé sur une modification de l'algorithme de conversion de voix déjà présentée.

Soit $S = \{c_1, ..., c_n\}$ une courte séquence de trames (fenêtres) consécutives de parole sélectionnés à partir d'un énoncé du locuteur ciblé. L'algorithme cherche une nouvelle séquence de trames de parole S^* qui maximise le score d'un système VAL donnée et donc le potentiel de spoofing.

Chaque trame $c(t)$ appartenant à S est d'abord transformé dans le domaine de fréquence avec conversion de voix où nous avons maintenant:

$$C'(f) = \frac{|H_c^*(f)|}{|H_c(f)|} C(f) \tag{B.3}$$

Le problème de l'identification de l'ensemble de filtres $H_S^* = \{H_{c_1}^*(f), ..., H_{c_n}^*(f)\}$ est formulé comme un problème d'optimisation. Au lieu d'estimer chaque filtre indépendamment en utilisant l'équation B.2, l'ensemble des filtres est optimisée conjointement en utilisant algorithmes génétiques. Tous les détails sont présentés dans [14].

Le système VAL utilisé pour la génération de signaux artificiels est le GMM-UBM sans normalisation de score[2] et avec la même configuration présentée dans la section B.2.1.2, mais en utilisant le NIST SRE'08 en lieu de NIST SRE'04 pour l'apprentissage de l'UBM.

Le signal de parole X est divisé en trames de 20 ms avec à décalage régulier de 10 ms. Les scores du VAL sont générés pour chaque trame afin d'identifier le court intervalle $T = \{c_1, \ldots, c_n\}$ dans X avec le meilleur score moyen. Nous avons mené des études avec des valeurs de n entre 1 et 20 cadres et observé de bons résultats avec une valeur de $n = 5$.

L'algorithme génétique a été mis en œuvre en utilisant l'outil MATLAB Global Optimization Toolbox V3.3.1. Sauf pour le nombre maximum de générations qui est fixé à 50, nous avons utilisé la configuration par défaut de MATLAB. Pour des informations détaillées sur la configuration du système, les lecteurs sont appelés a lire [12].

[2]Notez que cette configuration de système VAL est utilisé pour la génération de signaux artificiels, et peut ou ne peut différer de la configuration des systèmes VAL de base dans cette thèse

Synthèse vocale

Les attaques de type spoofing avec synthèse de la parole ont été mis en œuvre en utilisant le système de synthèse vocale (HTS) à base de modèle de Markov caché (HMM)[3] et l'approche spécifique décrit dans [204]. La paramétrisation comprennent caractéristiques STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum), coefficients Mel-cepstre et le logarithme de fréquence fondamentale (log F_0) ainsi que leurs coefficients delta et d'accélération. Caractéristiques spectrales acoustiques et les probabilités de durée sont modélisés en utilisant les modèles semi-Markov cachée de distribution multispace (MSD-HSMM) [171]. Modèles d'excitation, spectrales et la durée dépendantes du locuteur sont adaptées de modèles indépendants correspondant selon une stratégie d'adaptation au locuteur dénommé contrainte maximale structurelle une régression linéaire de posteriori (CSMAPLR) [203]. Enfin, les signaux de domaine de temps sont synthétisés en utilisant un vocodeur sur la base de filtres d'approximation du spectre Mel-logarithmique (MLSA). Ils correspondent à des coefficients Mel STRAIGHT-cepstraux et sont entraînés par un signal d'excitation mixte et formes d'onde reconstruits en utilisant le méthode *overlap-add* de chevauchement synchrone (PSOLA) [140].

B.2.3 Protocoles & métriques

B.2.3.1 Protocole pour les transactions biométriques licites

Sauf pour l'utilisation de la condition 8conv4w à la place de la condition 1conv4w, les protocoles pour les transactions biométriques licites sont exactement les mêmes que celles de tous les résultats de base mentionnés à la Section B.2.1.

Les nouveaux protocoles/conditions se traduisent par un nombre légèrement réduit de locuteurs et des enregistrements de base utilisé pour des expériences rapportés précédemment. Un résumé des tailles des bases de données illustrant le nombre de clients, les essais de clients et essais imposteurs est illustré dans le tableau B.1 pour les deux bases de données de développement et d'évaluation. Le base de développement contient 201 clients masculins alors que le base d'évaluation contient 298 clients masculins. Ces chiffres sont conformes au protocole de base et les trois protocoles de spoofing. Le protocole précis est

[3] http://hts.sp.nitech.ac.jp/

NIST SRE - 8conv4w-1conv4w - locuteurs masculins		
	Dév. (NIST'05)	Éval. (NIST'06)
Base	201/984/8962	298/1344/12648
Signaux Artificiels	201/984/201	298/1344/298
Conversion de la Voix	comme base	
Parole Synthétisée	comme base	

Table B.1: Taille des bases de données utilisées pour l'évaluation du spoofing. Les chiffres illustrent le nombre d'essais clients / essais de véritables clients / et essais d'un imposteur.

non-exhaustive et exactement tel que défini par le NIST.

B.2.3.2 Protocole pour des attaques de type spoofing

Les protocoles d'évaluation de le spoofing sont identiques à le protocole licites seulement que, pour chaque tentative d'accès imposteur, l'échantillon d'essai est remplacé par l'un des trois attaques déjà mentionnées.

Pour les bases de données NIST-SRE, toutes les données utilisées pour générer les attaques de type spoofing vient de l'une des sessions ne sont pas utilisés pour les tests. Du 8 disponibles, le premier est utilisé pour les tests, ainsi sessions 2-8 peuvent être utilisés pour générer les attaques. En outre, toutes les autres données appropriées, par exemple, le base de données NIST 2008, seront utilisés comme données de base indépendants (c.-à-d. pour apprendre un modèle UBM utilisé dans la conversion de la voix). À l'exception des modifications de segments imposteur par usurpation d'essai, les protocoles sont exactement les mêmes que celles décrites ci-dessus. En conséquence il n'y a pas de chevauchement entre les données utilisées pour former les modèles de clients et celui utilisé pour le spoofing.

B.2.4 Résultats

Nous avons couru une série d'expériences conçues pour comparer la performance de les systèmes GMM-UBM, GSL, GSL-NAP, GSL-FA, FA et IV-PLDA pour les locuteurs masculins. Nous avons également examiné trois attaques de type spoofing différents, y compris les signaux artificiels, la conversion de la voix et la parole synthétisée défini à la Section B.2.2.2.

Système	Development		Evaluation	
	no-norm	norm	no-norm	norm
GMM-UBM	8.2	8.1	9.1	8.6
SGL	7.8	7.8	7.9	8.1
SGL-NAP	5.9	5.9	6.3	6.3
SGL-FA	5.1	5.1	6.1	5.7
FA	4.7	5.1	5.6	5.6
IV-PLDA	4.2	4.3	3.4	3.2

Table B.2: Taux d'erreurs égales (EER %)pour les six systèmes VAL (locuteurs masculin). Les résultats sont illustrés pour la base de données le développement (NIST'05) et l'évaluation/test (NIST'06).

B.2.4.1 Baseline

Tous les systèmes ont été optimisés sur l'ensemble de développement et ont ensuite été appliquées sans modification à l'ensemble de l'évaluation en suivant les protocoles de la Section B.2.1. Les taux de change effectifs pour chaque système sont présentés dans le tableau B.2 et montrent l'évolution de la performance avec des approches différentes pour compenser la variation de l'intersession.

Pour le locuteurs masculins les plus performants du système IV-PLDA (jugé à partir de l'ensemble de développement) donne un EER de 3,2% sur l'ensemble de l'évaluation. Cela se compare bien au système GMM-UBM où l'EER respective est de 8,6%. Après comparaison de ces résultats à ceux rapportés dans les plus récentes campagnes NIST-SRE nous notons que les systèmes testés représentent l'état de l'art dans la technologie actuelle de reconnaissance automatique des enceintes et sont des candidats donc appropriés pour évaluer la menace de le spoofing et pour les tests de contre-mesure.

De l'analyse des les résultats dans le tableau B.2 nous observons que les différences de performance entre les systèmes sans normalisation de score ne sont pas significatifs. Il est à noter que les configurations des systèmes VAL restent les mêmes pour les expériences dans les sections suivantes et ne sont pas nécessairement les configurations des systèmes qui donnent la meilleure performance sur les protocoles de transaction licites.

B.2.4.2 Spoofing

Les résultats sont illustrés par le courbes DET et à travers un résumé des taux de fausse acceptation (FAR) pour le taux de faux rejets (FRR) fixée (une analyse plus détaillée est présentée dans la version anglaise de cette thèse).

Les expériences comprennent les attaques à haute effort et un exemple d'un attaque de faible effort avec bruit blanc.

Alors que dans cette section, nous présentons les résultats pour les six systèmes différents, nous nous concentrons sur les systèmes GMM-UBM et IV-PLDA. Le premier est sans doute l'approche la plus populaire pour la reconnaissance du locuteur tandis que le deuxième est représentatif de l'état de l'art et donne la meilleure performance parmi les six systèmes testés, selon les résultats de la Section B.2.4.1.

Attaques d'effort de haut niveau

Résultats pour les systèmes VAL avec et sans normalisation de score sont illustrés par moyen d'une collection de FAR présenté dans le tableau B.3(b) et le tableau B.3(a), respectivement. Les tableaux indiquent les résultats sur les performance des systèmes VAL pour les transactions licites et sous les attaques de spoofing.

Les FAR sont calculées en fixant le FRR de chaque système VAL à sa valeur de référence de l'EER (%). Résultats de la performance des six systèmes VAL pour les transactions licites, illustré dans la première colonne des tableaux B.3, sont donc directement comparables aux valeurs de EER présentés dans les tableaux B.2. Nous rapportons des performances similaires pour chacun des six systèmes VAL analysés, malgré les protocoles de formation venant de 1conv4w (section B.2.1) ou 8conv4w (section B.2.2). Le système IV-PLDA donne à nouveau la meilleure performance avec un FAR de 3%.

Depuis les attaques considérés, ceux avec voix convertis apparaissent comme la menace la plus grave pour la plupart des systèmes VAL; FAR base entre 3% et 9% et augmentation avec des valeurs comprises entre 71% et 94% pour tous les systèmes VAL analysés. Les petites augmentations, mais toujours significatifs dans le FAR sont signalés pour la synthèse vocale, avec des augmentations de fausses acceptations entre 36% à 87%.

Pour ces derniers, des dégradations significatives sont observées dans tous les cas, sauf pour les attaques de signaux artificiels et les trois systèmes GSL et

(a) Systèmes VAL sans normalisation de score.

Système/Attaque	-	AS	VC	SS	WN
GMM-UBM	9.1	93	78	87	53
SGL	7.9	2	92	41	2
SGL-NAP	6.3	8	88	55	2
SGL-FA	6.1	1	90	39	4
FA	5.6	73	71	77	36
IV-PLDA	3.0	11	94	54	13

(b) Systèmes VAL avec normalisation de score.

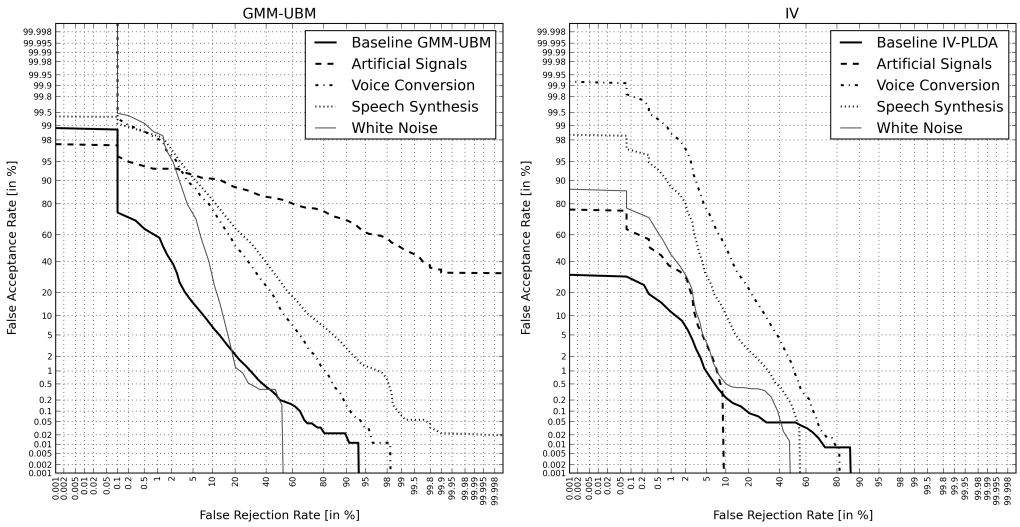
Système/Attaque	-	AS	VC	SS	WN
GMM-UBM	8.6	70	91	72	4
SGL	8.1	2	92	42	3
SGL-NAP	6.3	21	88	57	4
SGL-FA	5.7	19	73	56	5
FA	5.6	38	83	59	8
IV-PLDA	2.9	16	85	36	1

Table B.3: Taux de fausses acceptations (FAR %) pour des FRR fixés à les valeurs EER des six systèmes de base pour les attaques avec signaux artificiels (AS), conversion de la voix (VC), synthèse de la parole (SS) et bruit blanc (WN). La première colonne correspond à la performance VAL sans attaques (transactions licites).

IV-PLDA. En outre, les systèmes GMM-UBM et FA sont relativement plus robuste que les systèmes GSL aux attaques avec conversion de voix (étant FA le plus robuste des six systèmes), alors que l'inverse semble se produire avec des attaques de spoofing avec synthèse de la parole.

L'impact ambigu de la normalisation de score est également visible dans le tableau B.3(b). Pour le système IV-PLDA, la normalisation symétrique de score (S-norme) contribué à diminuer les valeurs FAR face à les attaques de type spoofing, par exemple, pour les attaques avec synthèse vocale FAR a diminué de 54% à 36%. En revanche, dans certains cas, par exemple, pour signaux artificiels ou synthèse de la parole et des systèmes GSL, le FAR a augmenté après l'application de la normalisation de score. L'influence de la normalisation de score dans le visage de le spoofing est toujours pas clair et l'objet de recherches plus poussées.

DET courbes sont présentés dans la Figure B.1 Les courbes sont illustrés pour le système GMM-UBM (a) et le système IV-PLDA (b) sans normalisation de



(a) Système GMM-UBM (b) Système IV-PLDA

Figure B.1: Performances de vérification de locuteur des systèmes GMM-UBM et IV-PLDA. Courbes indiquées pour l'expérience de base et les différentes attaques de spoofing.

score.

Par rapport à les synthèses VAL de référence, la conversion de la voix et de la parole à la fois provoquer des augmentations du taux de fausses acceptations (FAR) dans toute la gamme de seuils et la menace de la synthèse vocale est légèrement supérieure à celle de la conversion de la voix pour les systèmes GMM-UBM, et vice-versa pour les systèmes IV-PLDA. Les signaux artificiels provoquent également des augmentations dans le FAR pour le système GMM-UBM (pour lequel il a été optimisé). Pour les FRR inférieurs a 3%, les signaux artificiels donnent la plus forte augmentation des FAR. De la Figure B.1(b) nous notons que cette configuration particulière de signaux artificielles ne représente pas une menace pour les systèmes IV-PLDA, cependant, comme mentionné dans la Section 5.2.1.1 (version anglaise), du première études suggèrent que les signaux artificiels formés avec un système similaire (IV-PLDA) représentent une menace sérieuse.

Un exemple d'attaque de effort faible

Un exemple d'attaque d'effort faible est présenté dans cette thèse pour motiver d'autres recherches sur la menace hypothétique cette attaques.

Nous rapportons une petite expérience dans laquelle les enregistrements des imposteur sont remplacés par des signaux ne contenant que du bruit blanc et donc testés contre les six systèmes VAL. Résultats liés à cette expérience sont montré dans la quatrième colonne dans les Tableaux B.3 pour les six systèmes VAL et Figure B.1(a) et Figure B.1(b) pour IV-PLDA systèmes GMM-UBM et, respectivement.

Tableau B.3(a) montre que les approches à base de SGL-non sont en effet vulnérables aux attaques de bruit blanc, tandis que sur l'utilisation de la note normalisation des actes généraux comme un bon contre-mesure contre une telle menace. Ci-dessous TRF de 10% et 2% pour GMM-UBM (figure B.1(a)) et IV-PLDA (figure B.1(b)), blancs attaques de bruit donnent augmentation des FAR-dessus de 30%.

B.2.5 Discussion

L'approche de signaux artificiels proposés dans cette thèse est une attaque dépendant du système c.-à-d. l'efficacité de l'attaque dépend sur les similitudes entre le système VAL ciblée et le système VAL utilisé pour synthétiser l'attaque. Cependant, les résultats présentés dans le Tableau B.3 montrent que des signaux artificiels provoquent des augmentations significatives sur le FAR d'une variété de systèmes VAL, spécialement pour ceux avec des scores normalisés, même pensé qu'ils ont été générés uniquement en utilisant un système GMM-UBM sans normalisation de score. En outre, notre travail rapporté dans [12] montre que les signaux artificiels sont également efficaces lorsque la configuration du système VAL utilisée pour le spoofing différé de la configuration des systèmes VAL cibles. Les travaux futurs devraient évaluer le risque d'attaques avec des signaux artificiels qui sont synthétisées en utilisant des systèmes plus sophistiqués c.-à-d. IV-PLDA.

La normalisation de score joue un rôle ambigu dans le visage de le spoofing. Les résultats présentés dans la Section B.2.4.2 montrent des différences significatives dans les FAR des systèmes VAL avec et sans normalisation de score lors d'un essai contre le même attaque, mais si la normalisation de score joue en faveur ou contre le spoofing dépend des attaques et des systèmes VAL considéré. Toutefois, nous notons que la normalisation de score peut fournir robustesse contre les attaques spécifiques (la normalisation de score provoque des taux plus bas pour bruit blanc) et les systèmes spécifiques (idem pour IV-PLDA), étant celui-ci une observation intéressante depuis la normalisation de score est pas prise en compte systèmes i-vecteur. Dans tous les cas, de plus amples recherches sont nécessaires dans ce sens.

Enfin, même pensé il n'y a pas des travaux antérieurs sur des attaques par spoofing généralisées, des déductions peuvent être faites sur la base des observations de travail expérimental de différentes sources. Résultats de la Section B.2.4.2 suggèrent que les signaux audio contenant un bruit blanc pur peuvent être légèrement une meilleure attaque que imposteur naïf, même si l'augmentation des fausses acceptations n'est pas significative, nous pensons que ces résultats suffisent à encourager davantage recherche.

B.3 Évaluation des contre-mesures

Les expériences de base et de spoofing ont été signalés dans le Section B.2. Comme il y'a de nombreuses approches de la reconnaissance du locuteur dans la littérature, les deux ont été étudiés avec une gamme de systèmes c.-à-d. systèmes GMM-UBM, GSL, GSL-NAP, GSL-FA, FA et IV-PLDA. Nous avons également examiné trois attaques de type spoofing d'haute effort comme les signaux artificiels, la conversion de la voix et de la parole synthétisée et aussi un exemple d'une attaque d'effort faible avec bruit blanc qui n'est pas abordée dans ce section.

Dans la suite, nous évaluons la performance des trois contre-mesures. Les évaluations comprennent l'évaluation de la performance en fonctionnement autonome et aussi en opération avec les systèmes VAL de base préalablement définis pour cette thèse. Les contre-mesures (avec des spécifications et configuration) sont décrites à la Section B.3.1, les résultats sont présentés dans la Section B.3.2 et sont discutés dans la Section B.3.3

B.3.1 Spécifications pour contre-mesures

Trois approches différentes sont envisagées. Ils sont la contre-mesure sur la base de la détection de motif répétitive (RPD) développé pour les signaux artificiels, la contre-mesure sur la base du analyse de distances de paires (PWD) pour la protection contre les attaques à la conversion de la voix et la contre-mesure sur la base des motifs binaires locaux (LBP) pour la protection généralisée.

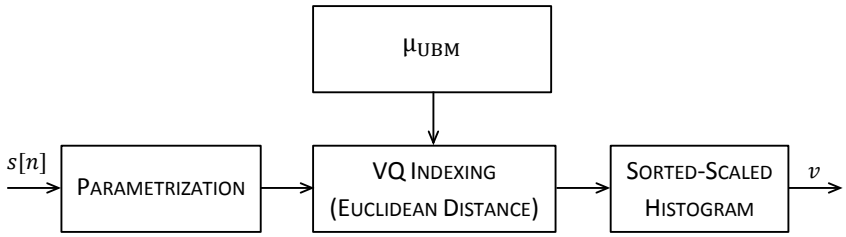


Figure B.2: Schéma de génération du vecteur caractéristique au niveau de l'énoncé (v).

B.3.1.1 Contre-mesures RPD

Étant donné un énoncé de parole, caractéristiques de plus haute niveau peuvent être extraites au niveau de cadre, niveau du mot, le niveau de phrase ou au niveau de l'énoncé. Dans ce travail, nous étudions une caractéristique aux niveau de l'énoncé qui peut être rapidement calculé à partir des caractéristiques aux niveau de cadre.

La Figure B.2 montre un schéma de notre approche pour calculer les caractéristiques de niveau de l'énoncé. Les paramètres classiques extraites du signal d'entrée sont indexés en utilisant quantification vectorielle (VQ) où les moyens de l'UBM agit comme le dictionnaire de codes. L'histogramme résultant est réorganisé sur la base des fréquences d'occurrence comme dans un diagramme de Pareto et les fréquences sont mis à l'échelle par rapport au premier composant pour obtenir une nouvelle vecteur caractéristique v.

Pour un énoncé réel de la parole, le vecteur caractéristique v aura une distribution exponentielle lisse alors que les signaux artificiels présenteront une distribution avec un pic dominant dans le premier coefficient. Cela facilite la classification robuste entre les signaux artificieux et les parole réelle.

Une contre-mesure pour détecter des signaux artificiels utilisant des vecteur caractéristique sur la base du détection de motif répétitive (RPD) est mis en œuvre par moyen d'un simple, *classificateur à distance moyenne*. Ce classificateur est basé sur une mesure de similarité entre le vecteur v_t et le vecteur moyenne v_{mean} obtenu à partir d'exemples de formation d'énoncés vocaux authentiques. Pour cette contre-mesure la distance cosinus a montré de bons résultats et, par conséquent, il a été choisi comme mesure de similarité. Une mesure de similarité cosinus élevé indique une probabilité plus élevée de spoofing.

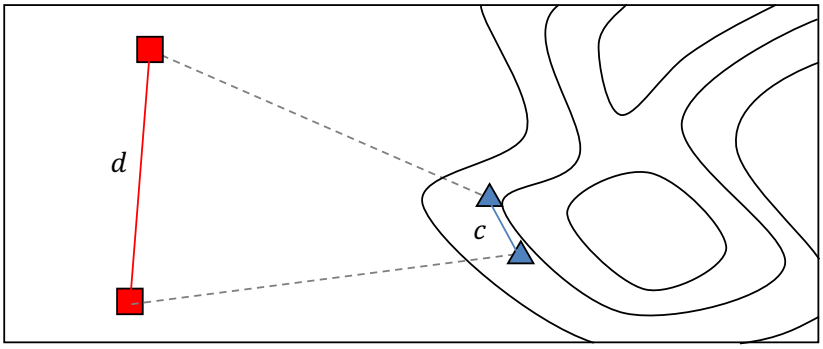


Figure B.3: Une illustration de la conversion de la voix dans l'espace des caractéristiques montrant le déplacement de deux vecteurs consécutifs vers un maxima local commun. Nous attendons généralement $c < d$.

B.3.1.2 Contre-mesure PWD

L'approche de conversion ce voix signalée dans le Section B.2.2.2 décale la pente spectrale d'un usurpateur vers celle d'un locuteur cible, selon l'équation B.2. Notre contre-mesure exploite le changement attendu des vecteurs de caractéristiques des trames consécutives vers le même, plus proche des maxima locaux de la fonction de vraisemblance d'un modèle GMM d'un locuteur cible en particulière. Ce principe est illustré dans la Figure B.3 pour deux vecteurs de caractéristiques consécutives dans un espace à deux dimensions. Dans ces conditions, la distance relative entre vecteurs consécutifs (carrés rouges) sera réduite (triangles bleus). En conséquence, nous avons étudié une nouvelle contre-mesure de détecter ce phénomène afin de distinguer entre la voix converti et un signal de parole authentique.

Nous avons effectué des expériences initiales pour valider ce phénomène dans le espace de caractéristiques (normalisée) asr, l'espace LPCC et dans l'espace LPC. Les distances par paires dans l'espace LPC présentent les plus grandes différences entre la parole et la voix véritable converti, donc ils sont utilisés dans toutes les expériences ultérieures rapportés ici.

La contre-mesure est dépendante du locuteur et exploite les différences dans la distribution des distances les paires entre la signal de voix de test $s[n]$ et la signal de voix d'apprentissage (celui utilisé pour former le modèle GMM du système VAL). Le pourcentage de chevauchement entre les deux distributions forme un score qui est ensuite seuillée de classer $s[n]$ comme voix réel ou spoofing. Lorsque les deux distributions sont normalisés, le pourcentage de chevauchement est comprise entre zéro et l'unité. Un score haut indique une

voix réel alors que un score bas indique voix converti.

Comme dans d'autres travaux antérieurs [54, 200], la contre-mesure proposée est intégré avec le système VAL comme une étape indépendante (post-traitement). La contre-mesure fonctionne sur les vecteurs LPC d'ordre 19^{th} recalculées à partir de un signal de domaine temporel $s[n]$. Le fenêtrage est le même que pour les systèmes VAL et de conversion de voix (bien que différentes longueurs de trame fournissent des résultats similaires). Nous prenons en compte que les trames déterminés pour contenir la parole vocale. Les trames voisée ont été détectée en utilisant l'algorithme RAPT [183] dans l'outil VOICEBOX[4] avec une configuration par défaut.

B.3.1.3 Contre-mesure LBP

Sur la base de nos travaux précédents [6], nous hypothéquons que la parole réel peut être distinguée de le spoofing en fonction des différences dans la 'texture' spectro-temporelle. La motivation découle de l'hypothèse que normalement les représentations spectrales de niveau supérieur sont plus difficile de synthétiser que l'information au niveaux de trame. La nouvelle contre-mesure dans ce document est basée sur l'application d'une approche standard pour l'analyse de texture connu comme motifs binaires locaux (MBL ou LBP en anglais) [147].

La Figure B.4 indique comme l'analyse de LBP est appliqué à un 'image' de deux dimensions d'un énoncé de parole, où ici l'image est une cepstrogramme formé de la concaténation des éléments cepstraux traditionnels, y compris la vitesse et l'accélération. Les LBPs sont déterminés pour chaque pixel, résultant ainsi en une nouvelle matrice de la gamme dynamique réduite, ici appelé un 'textrogram'. Les dimensions de la textrogram sont déterminées par le nombre de composants de chaque vecteur de caractéristique et la durée du signal de parole. Le textrogram capture information spectro-temporelle au-delà de paramétrages dynamiques classiques.

Classification des énoncés de parole que soit la parole authentique ou falsifiée discours est basé sur un nouvel ensemble de caractéristiques extraites de la textrogram. Un histogramme de LBP est construit pour chaque ligne de la textrogram. L'ensemble des histogrammes sont normalisé et le vecteur anti-spoofing est formé de la concaténation de chaque histogramme en un seul super-vecteur.

[4] `http://http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`

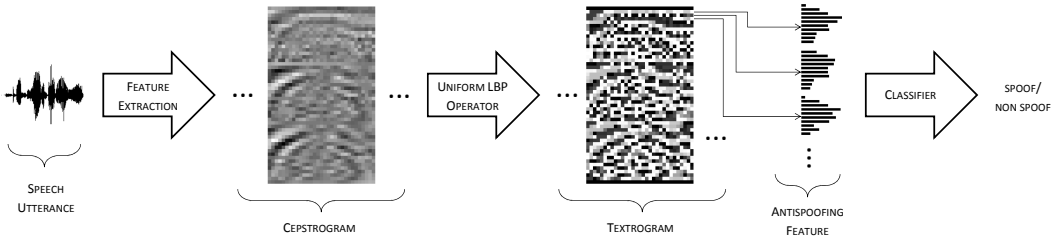


Figure B.4: Application de l'analyse de LBP uniforme à une cepstrogramme pour obtenir le textrogram. Les motifs non uniformes sont jetés et les histogrammes restantes sont normalisées et enchaînés pour former un vecteur de anti-spoofing.

L'analyse de LBP est appliqué à cepstrogrammes composés de 51 coefficients: 16 LFCC et de l'énergie ainsi que leurs coefficients delta et delta-delta. Le fenêtrage est le même que pour les systèmes VAL. Nous prenons en compte que les cadres déterminés pour contenir la parole, c.-à-d. ceux également utilisé pour VAL.

Nous avons effectué des expériences avec opérateurs $LBP_{4,1}$, $LBP_{8,1}$, $LBP_{8,2}$, and $LBP_{16,2}$ et leurs versions uniformes utilisant la LBP la disposition du public MATLAB mise en œuvre de l'Université de Oulu[5]. Nos meilleurs résultats ont été obtenus avec une $LBP^{u2}_{8,1}$ opérateur ne considérant que les 58 possibles modèles uniformes. Histogrammes sont créés pour tous, mais les premières et dernières lignes de la textrogram, obtenant ainsi un $58 \times (51 - 2) = 2842$ fonction de la longueur vectorielles.

Nous évaluons trois classificateurs différents. Dans tous les cas, les attaques avec la synthèse de la parole et des signaux artificiels représenter l'univers des attaques inconnues. Pour les deux premières approches d'une-classe, seule la voix convertie a été utilisé pour l'optimisation. Le premier approche est d'une-classe[6] et dépendant du locuteur, selon laquelle les scores scores sont obtenus par moyen de la comparaison du vecteur de LBP extrait de l'énoncé de test et de l'énoncé de apprentissage en utilisant un noyau d'intersection des histogramme. Le deuxième est aussi un classificateur SVM d'une-classe (approche indépendante du locuteur) et le troisième est un classificateur SVM à deux classes. Tous les classificateurs SVM sont implémentées en utilisant la bibliothèque LIBSVM [44][7] et sont à l'écoute en utilisant seulement une des parole réels ou voix convertis dans le jeu de développement.

[5] `http://www.cse.oulu.fi/CMV/Downloads/LBPMatlab`
[6] Seulement discours réel est utilisé pour la modélisation.
[7] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

B.3.1.4 Protocols & métriques

Le travail dans cette thèse liée aux évaluations de contre-mesures a été menée selon certaines lignes directrices générales. D'abord, chaque contre-mesure proposée est évaluée en fonctionnement autonome contre les trois attaques de type spoofing et avec la même base de données de spoofing et protocoles définis à la section B.2.2. Par conséquent, chaque contre-mesure proposée est évaluée avec les six systèmes de VAL déjà définies.

En plus des évaluations de contre-mesures indépendantes du système, les résultats sont présentés à travers un ensemble de trois courbes DET contenant chacune quatre profils (cette analyse ne est disponible que dans la version anglaise de cette thèse). L'évaluation des contre-mesures est soumis à deux types d'erreurs notée 'False Living Rate' (FLR) et 'False Fake Rejection' (FFR), qui sont similaires aux FAR et FRR, respectivement. Le point auquel FLR = FFR est appelée Classification d'Erreur moyenne (ACE), qui est similaire au EER.

B.3.2 Résultats

Les travaux expérimentaux liés à la détection de motif répétitif (RPD), l'analyse de distances par paire (PWD) et les motifs binaires locaux (LBP) à partir des contre-mesures sont signalés dans la Section B.3.2.1, Section B.3.2.2 et la Section B.3.2.3, respectivement. Alors que travaux expérimentaux de base et de spoofing sont évalués pour les systèmes VAL avec et sans normalisation de score, l'évaluation des contre-mesures sont présentés uniquement pour ce dernier.

B.3.2.1 Détecteur de motif répétitif

La contre-mesure basée sur l'analyse de la distribution des vecteurs de caractéristiques des systèmes VAL a été conçu pour détecter des motifs répétitifs qui font des attaques par spoofing. Des améliorations significatives sur la performance de base ont donc été prévu dans ce cas de signaux artificiels. La Figure B.5(a) montre une courbe DET qui vise à évaluer la performance de la contre-mesure indépendamment du système VAL. Comme prévu, la contre-mesure est extrêmement efficace dans le cas de signaux artificiels et détecte toutes les attaques de spoofing. En revanche, toutefois, l'ACE, où le RPF et la FFR ont la même valeur, est près de 40% pour la conversion de voix et 20%

pour la synthèse vocale.

B.3.2.2 Analyse de distances par paires

Ce travail rapporte les résultats expérimentaux sur la nouvelle contre-mesure qui exploite la réduction des distances des vecteurs caractéristiques consécutives.

Cette contre-mesure a été spécialement conçu pour résoudre le problème de la conversion de la voix. Des améliorations significatives sur la performance de base ont donc été prévu dans ce cas.

Une courbe DET qui vise à évaluer la performance de la contre-mesure indépendamment du système VAL est présentée dans la Figure B.5(b). Comme prévu, la contre-mesure est très efficace dans le cas de la conversion de la voix et détecte la plupart des attaques de type spoofing (ACE de 2,5%). En revanche, cependant, l'ACE est de 35% pour les signaux artificiels et 10% de la parole synthétisée.

B.3.2.3 Contre-mesure LBP

Nous avons introduit un nouveaux vecteur de caractéristiques pour la détection de le spoofing basée sur l'analyse du paramétrage de la parole conventionnelle utilisant des motifs binaires locaux (LBP). Cette vecteur, avec ce qui est, au mieux de notre connaissance, la première approche de classification d'une-classe à la détection de spoofing, résultats dans une contre-mesure généralisé pour la protection des systèmes VAL.

Résultats pour la nouvelle contre-mesure avec des vecteurs LBP et chacun des trois classificateurs déjà proposes sont illustrés dans le Tableau B.4. Pour la SVM d'une-classe, nous avons obtenu nos meilleurs résultats avec une fonction radiale de base du noyau, tandis qu'un noyau linéaire a donné de meilleurs résultats pour le classificateur deux-classes. Comme prévu, par rapport aux classificateurs d'une-classe, le classificateur bi-classe offre les meilleures performances pour la condition sur laquelle il est optimisée (de conversion de voix). Voici l'ACE est 0%. Cependant, pour les deux attaques de type spoofing pas vu lors de l'optimisation, la performance est médiocre. Alors que les classificateurs d'une classe ne réussissent pas aussi bien que le classificateur à deux classes pour les attaques de conversion de voix , ACE de 8% et 5% ne sont que marginalement plus élevé que le ACE de référence de 3%. Plus important encore, les classificateurs d'une classe sont considérés bien généraliser

Contre-mesure		Attaque		
Feature + Classifier	Speaker	VC	SS	AS
RPD + 1-class MDC [12]	independent	40	20	**0**
PWD + 1-class OIC [6]	dependent	**3**	10	35
LBP + 1-class HIK [11]	dependent	**8**	1	0
LBP + 1-class SVM [5]	independent	**5**	0.1	0
LBP + 2-class SVM [5]	independent	**0**	56	25

Table B.4: Performances de contre-mesures en termes de ACE (%) pour chacun des cinq contre-mesures présentées dans cette thèse et les attaques avec conversion de voix (VC), la synthèse de la parole (SS) et des signaux artificiels (AS). Les valeurs en gras correspondent aux attaques pour lesquels la contre-mesure a été optimisés

à la parole synthétisée et signaux artificiels. Voici les ACE sont tous inférieurs à 1%.

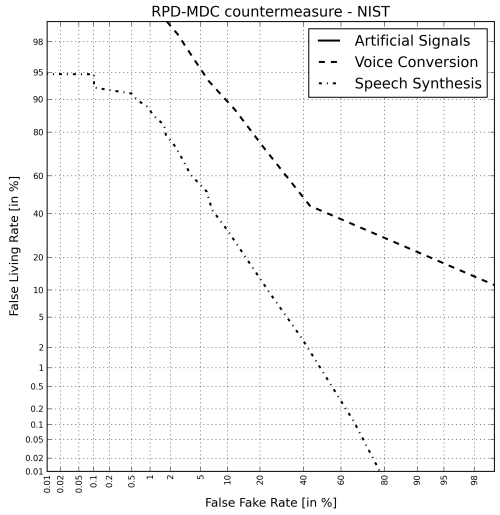
Le Tableau B.4 montre que la meilleure performance globale est obtenu avec le SVM d'une classe. Des courbes DET qui montrent la performance de la contre-mesure indépendamment des systèmes VAL sont illustrées à la Figure B.5(c).

B.3.3 Discussion

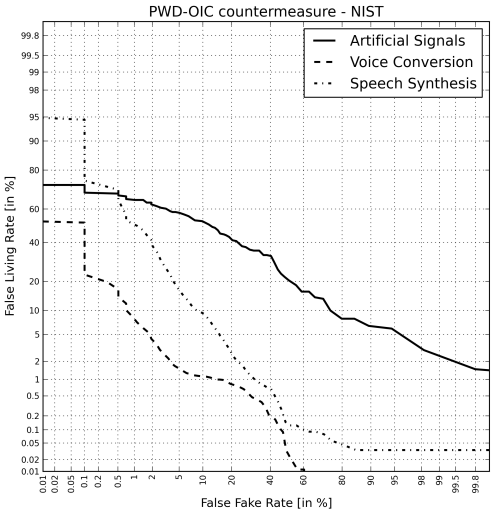
Le Tableau B.4 résume les résultats liés à la performance des différentes contre-mesures présentées dans cette thèse. Ils sont le détecteur de motif répétitif (RPD), combinées à classificateurs de distances moyenne (MDC), l'analyses des distance par paires (PWD) combinés avec des classificateurs de l'indice de chevauchement (OCI) et des motif binaire locaux (LBP) combinés avec des classificateurs d'intersection des histogrammes (HIC) et également avec des classificateurs SVM de une et deux classes.

La contre-mesure PWD a été développé pour détecter voix convertis et sont basées sur la contraction de l'ensemble de caractéristiques qui se produit en conséquence du processus de conversion. Comme prévu, la contre-mesure est largement efficace pour détecter les attaques d'usurpation avec conversion de voix. Il est encourageant, même lorsqu'il est testé contre d'autres formes d'attaque par usurpation pour lesquels il n'est pas optimisé, par exemple parole synthétisée, la contre-PWD fournit toujours une performance acceptable dans certains cas.

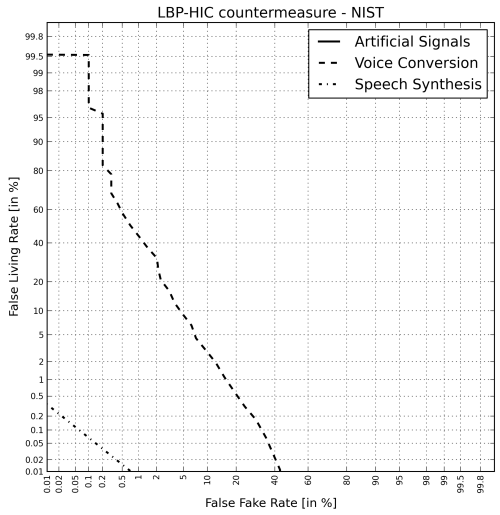
Les contre-mesures avec l'analyse de texture montrent des performances en-



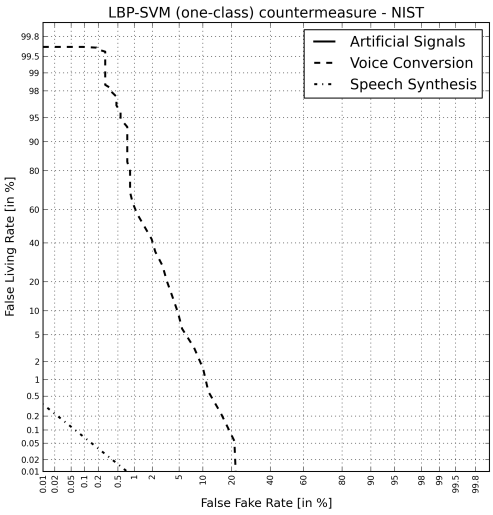
(a) Contre-mesure RPD + MDC



(b) Contre-mesure PWD + OIC



(c) Contre-mesure LBP + HIK



(d) Contre-mesure LBP + SVM 1-classe

Figure B.5: Courbes DET pour les contre-mesures RPD avec classificateur MDC, PWD avec classificateur OIC et LBP avec classificateurs HIC et SVM d'une classe évalués indépendamment du système VAL.

courageantes, donnant des performances de détection presque parfaite des signaux artificiels et synthèse de la parole, et aussi bonne performance pour la conversion de la voix. Nous observons que la combinaison appropriée de LBP avec classificateurs d'une classe et bi-classe permettrait de résoudre le problème de le spoofing pour les attaques et la configuration considérée dans cette thèse. Ce n'est cependant objet de nouvelles recherches.

D'autre part, des expériences initiales sur des signaux vocaux d'une durée considérablement réduits ont montré que l'efficacité des contre-mesures sur la base des vecteurs LBP dépend de la longueur du signal d'entrée. Cela peut se expliquer en observant qu'un petit nombre d'élocution ou caractéristiques sont insuffisantes pour générer des histogrammes représentatifs[8] de la même manière qu'un petit nombre de caractéristiques de parole sont insuffisantes pour générer un modèle GMM représentant [179]. Néanmoins, cela pose une limitation pour l'utilisation pratique des contre-mesures sur la base de LBP et est objet de nouvelles recherches.

B.4 Conclusions et perspectives d'avenir

L'état-de-l'art dans la vérification automatique du locuteur (VAL) indépendante du texte a progressé rapidement au cours des dernières années. Étonnamment, cependant, il y'a eu relativement peu de travail dans le développement de contre-mesures pour protéger ces systèmes VAL de la menace de spoofing.

Cette thèse contribue au développement d'une nouvelle génération de systèmes VAL plus sûres capables de fournir une reconnaissance fiable. Cette thèse analyse les vulnérabilités des systèmes VAL et les mécanismes utilisés par les attaques connues pour les tromper. Il recherche également de nouvelles menaces et introduit des nouvelles contre-mesures qui atténuent les effets de ces menaces. Plus important encore, il reconnaît certaines faiblesses fondamentales dans le développement de contre-mesures, telles que l'utilisation abusive de connaissance préalable, et établit certaines des bases pour des recherches plus poussées, telles que l'utilisation de systèmes de classeurs multiple (MCS) dans l'évaluation avenir avec généralisée contre-mesures.

En raison de la nouveauté de ce domaine de recherche, il y'a encore un long chemin à parcourir. Sans aucun doute, le principal problème concerne le manque de base de donnes standards de grande envergure, des protocoles ou des mesures qui pourraient être utilisés pour mener des évaluations sur le spoofing et de contre-mesures dans un sens plus juste et des résultats plus comparables, les normes qui sont nécessaires d'être définis à l'avenir. Autres conclusions et orientations futures de la recherche sont décrits ci-après.

[8] nous restons le lecteur que la fonction de LBP est essentiellement formé par un enchaînement d'histogrammes

B.4.1 Attaques de type spoofing

La littérature sur le spoofing est limitée à quatre attaques différentes, y compris des attaques par spoofing classiques telles que les attaques "replay" ou la mimique ainsi que les attaques les plus avancées telles que la conversion de la voix et la synthèse de la parole, tous montrés à provoquer une augmentation significative du taux de fausse acceptation de pointe de la technologie l'art des systèmes VAL.

Dans cette thèse, nous nous intéressons à l'analyse et la détection des attaques de spoofing *réussi*. Dans le cadre de cette thèse, si un signal de spoofing surmonte le module la détection de locuteur **et** le module de reconnaissance du locuteur d'un système VAL donné, alors le signal est une attaque de spoofing réussie (Section 4.1 version anglaise).

Cette thèse définit des attaques par spoofing par moyen d'une condition suffisante. Les menaces envisagées par cette large sens, la définition conservatrice de le spoofing comprennent, mais ne se limite pas aux attaques mentionnées ci-dessus. Pour valider cette approche cette thèse aborde les signaux de non-parole (signaux artificiels).

Les signaux artificiels constituent ainsi une menace sérieuse pour la fiabilité des systèmes VAL. Pour les systèmes et protocoles testés, tandis que la conversion de la voix et de synthèse vocales attaques ne sont pas toujours aussi efficaces que les signaux artificiels (selon le point de fonctionnement et le système VAL utilisé pour la formation de signal artificiel) la menace est néanmoins significatif. Si les signaux artificiels produisent des signaux non vocaux comme, conversion de voix et la synthèse vocale ont l'avantage de produire des signaux de parole comme. En ligne avec les précédents, les travaux connexes, cette contribution souligne l'importance des efforts visant à développer des contre-mesures dédiées, certains d'entre eux trivial, pour protéger les systèmes VAL de le spoofing.

Les résultats présentés ci-dessus montrent que tous les systèmes testés sont vulnérables à le spoofing par des attaques proposées qui provoquent des dégradations significatives de performance. Nous notons, toutefois, que les résultats présentés ci-dessus sont strictement liés à des techniques spécifiques, des systèmes et des protocoles qui représentent environ, mais certainement pas toutes les vulnérabilités du système VAL. Une étude complète de chaque attaque de spoofing représente un important projet de recherche en soi, c'est à dire plusieurs approches pour exprimer la conversion et la synthèse de la parole sont rapportés dans la littérature et il y'a sans aucun doute beaucoup

plus d'approches pour générer des signaux artificiels spécifiques spoofing que nous ne pouvons pas aborder dans le cadre de cette thèse. Par conséquent, tout ce travail démontre la vulnérabilité des systèmes à la pointe de l'art ce n'est que la première étape vers une compréhension complète de la menace de spoofing.

Enfin, cette thèse aborde également le spoofing en termes d'accessibilité (effort). Sauf pour les attaques "replay", toutes les attaques adressées ne sont guère accessibles pour la masse, soit en raison de la nécessité de compétences spécifiques (mimique) ou en raison de la nécessité d'une expertise spécifique (conversion de voix, la synthèse de la parole et des signaux artificiels). D'autres recherches devraient se concentrer dans déterminer si oui ou non les systèmes VAL peuvent être falsifiés avec d'autres signaux facilement générés. Ces telles attaques peuvent être plus représentatif du scénario pratique de spoofing.

B.4.2 Contre-mesures et intégration

Nombreuses études de vulnérabilité suggèrent un besoin urgent de d'adresse du spoofing et la solution ne semble pas être trivial. Par exemple, la première contre-mesure proposé pour cette thèse on a émis l'hypothèse qu'une simple routine d'évaluation de qualité de la parole, par exemple, peut être utilisé pour distinguer des signaux artificiels de signaux de parole d'origine. Les travaux rapportés dans [12] montre la première étude qui a utilisé les spécifications de l'UIT-T comme une contre-mesure basée en l'évaluation de la qualité de la parole, dans ce cas, contre les attaques avec des signaux artificiels. Contre l'intuition, les résultats ne sont pas satisfaisants, qui souligne la nécessité d'un effort soutenu pour le développement de contre-mesures.

Cette thèse rapporte trois nouvelles contre-mesures pour la protection des systèmes VAL contre le spoofing. Les deux premiers sont une approche trivial basé sur la détection de motif répétitif (RPD) et une approche basée sur l'analyse des distances des paires (PWD) pour détecter les attaques de type spoofing avec des approches spécifiques aux signaux artificiels et conversion de voix, respectivement. Alors qu'en général, les six systèmes VAL testés montrent vulnérabilités considérables contre ces deux attaques de type spoofing, ces deux contre-mesures sont présentés pour être cohérente et extrêmement efficace dans la détection d'attaques de spoofing pour lesquels ils ont été optimisés.

Pour fournir les contre-mesures avec une certaine souplesse à l'égard des at-

taques similaires (c.-à-d. mêmes attaques pour lesquelles les contre-mesures ont été conçus mais généré avec des configurations de type spoofing différents que le utilisés en laboratoire), chaque contre-mesure comprend un classificateur d'une classe formés uniquement avec des enregistrements de voix réel. Par conséquent, lorsqu'il est testé contre d'autres formes d'attaque de type spoofing pour lesquelles ils ne sont pas optimisés, les deux contre-mesures fournissent toujours de bonnes performances dans certains cas. Cependant, le résultat global n'est pas satisfaisante. Bien que chacune de ces contre-mesures est réussi à surmonter l'attaque spécifique considéré, dans les concepteurs de systèmes de réalité et les développeurs de contre-mesures ne peuvent pas assumer une telle connaissance préalable. Dans la pratique l'attaque de spoofing peut jamais être connu et ensuite la performance des contre-mesures existantes dans les scénarios pratiques n'est pas garantie.

En conséquence, il existe un besoin pour des contre-mesures généralisées ayant le potentiel pour détecter les attaques pour lesquelles ils ne ont pas été optimisés. Cette thèse traite de cette question dans une certaine mesure. La troisième contre-mesure proposée est basée sur les caractéristiques obtenues après l'analise des motif binaire locaux (LBP) des séquences de vecteurs acoustiques. Le vecteur de caractéristiques résultant, lorsqu'il est combiné avec classificateurs d'une classe est, au mieux de notre connaissance, la première approche généralisée de détection de le spoofing pour les systèmes VAL.

Les résultats montrent que la contre-mesure basée en LBP est moins efficace que des solutions spécifiques pour des attaques basés en conversion de la voix, mais que une détection presque parfaite est obtenue pour des attaques inédites qui provoquent des augmentations significatives de fausse acceptation. Étant moins dépendants des connaissances préalables, les points de travail à la possibilité de contre-mesures généralisées à plus grande valeur pratique.

Les résultats suggèrent également que les travaux futurs devraient envisager la combinaison de contre-mesures spécifiques et généralisées avec les systèmes de reconnaissance q'ils visent à protéger (ce dernier en raison du fait que, dans la pratique, une contre-mesure ne sera jamais utilisé de manière autonome). La combinaison de systèmes biométriques et de contre-mesures est donc un point clé dans la recherche future.

La combinaison de classificateurs, connus comme 'systèmes a multiple classeurs' (SMC, ou MCS en anglais), est également abordée dans cette thèse. Bien que, dans cette thèse, nous n'effectuons des recherches sur les techniques de fusion pour combiner les systèmes VAL et de contre-mesures, dans la Section 6.2 (version anglaise) nous introduisons une partie de la base de concevoir MCS

dans le contexte de le spoofing.

Une partie de la contribution de cette thèse est liée à la description des différentes approches de formuler le problème de la vérification du locuteur fiable (voir la Section 6.1 (version anglaise)). Aussi fondamental que cette tâche semble, il reste toujours sans consensus entre les recherches dans la communauté biométrique. Approches possibles pour la formulation du problème comprend des exemples provenant de l'approche holistique qui traite des usurpation au problème à deux classes classique à l'approche réductionniste qui assigne une classe par attaque.

En particulier, dans cette thèse le problème du spoofing et des contre-mesures est formulé comme un problème multi-classes avec détection des valeurs aberrantes (Section 6.1.3.3 version anglaise). Les évaluations dans cette thèse sont rapportés par la combinaison d'un maximum de deux classificateurs. Dans cette perspective, la combinaison d'un système de une contre-mesure généralisée avec un système VAL, qui donne les meilleurs résultats globaux de cette thèse, peut être considérée comme un problème de bi-classe avec détection de valeurs aberrantes.

Se adressant le spoofing et les contre-mesures comme un problème multi-classes avec détection des valeurs aberrantes apparaît comme l'approche la plus appropriée compte tenu de l'état de l'art dans le domaine. Cependant, pour motiver la recherche dans ce sens il y'a un besoin évident des évaluations formelles de spoofing et de contre-mesures. Les évaluations formelles avec de corpus, de protocoles et de mesures sont donc nécessaires pour stimuler la recherche de contre-mesures contre le spoofing dans les paramètres correctement contrôlées réfléchissantes des scénarios pratiques avec des attaques méconnues et variées. Cette discussion est abordée dans la section suivante.

B.4.3 Évaluations & bases de données

Même se ils proviennent de l'adaptation des bases de données standard, tout le travail passé a été effectuée sur des bases de données non standard des signaux de spoofing. Cela a souvent entraîné le développement d'un seul ou un petit nombre d'algorithmes de spoofing spécifiques pour les essais de spoofing générés. Les évaluations de contre-mesures sont donc biaisés vers les attaques spécifiques et manquent de généralité à de nouveaux algorithmes de spoofing ou entièrement nouvelles formes d'attaque qui sera probablement émerger dans l'avenir. La Figure 6.3 montre que les évaluations appropriées

jouent un rôle essentiel dans le cycle de conception du MCS. En ce sens, nous notons que les bases de données et des évaluations actuelles ne sont pas représentatifs des scénarios pratiques.

Afin de répondre à l'utilisation inappropriée des connaissances préalables dans les travaux futurs, il sera nécessaire de recueillir et de mettre à disposition des bases de données standard pour les évaluations de base et de spoofing, avec autant de variations que possible afin d'éviter les biais et plus-raccord. Bases de données standard seront alors encourager l'intégration de la détection des valeurs aberrantes (contre-mesures généralisées).

La conception de bases de données de type spoofing en adaptant bases de données standard, bien que si l'on préfère par rapport aux études à petite échelle impliquant des bases de données à des fins recueillies, peut également ne pas être d'exclusion de certains cas d'utilisation pratiques. Tel que discuté à la Section 4.3.2, avec cette configuration attaques sont simulés par spoofing au niveaux de post-capteur (niveaux de transmission). Cette configuration peut être acceptable dans le cas des applications de téléphonie, ou si le capteur, le canal et l'attaque sont toutes transformées linéaires, mais en réalité, c'est peu probable. La configuration est aussi irréaliste dans le cas de scénarios d'accès physiques où le microphone est fixé; les données de SRE, par exemple, contient divers canaux microphone et effets.

La principale raison pourquoi les attaques au niveau de la transmission sont beaucoup plus étudiés que la contrepartie du niveau du capteur liée à la relative facilité de générer les ensembles de données de type spoofing. Par exemple, pour simuler des attaques au niveau du capteur, les énoncés de spoofing peuvent être soit ré-enregistrées conformément aux protocoles spécifiques ou soit les énoncés de spoofing peut être contournés artificiellement avec différentes réponses impulsionnelles. Cette dernière approche apparaît beaucoup moins gênant pour mettre en œuvre, même si sa validité doit être étudiée. D'autres travaux devraient également étudier la relation entre les évaluations au niveau de transmission et le capteur (par exemple, peuvent évaluations réalisées à l'aide du niveau de la transmission de déduire des résultats de l'évaluation au niveau du capteur?), la perspicacité dans ce domaine seraient considérablement faciliter la conception de bases de données de spoofing.

En outre, la majorité des travaux antérieurs a également été menée dans des conditions adaptées, à savoir les données utilisées pour apprendre les modèles de locuteur cibles et que utilisés pour effectuer le spoofing ont été recueillis dans la même ou similaire environnement acoustique et sur le canal identique ou similaire, alors que ce ne pourrait pas être réaliste. Afin de réduire le biais

dans les résultats générés en fonction de ces configurations, les travaux futurs devraient étudier l'impact pratique des différences entre les deux montages expérimentaux illustrés à la Figure 4.4. De préférence, les travaux futurs devraient inclure la collecte de nouvelles bases de données qui représentent plus fidèlement des scénarios pratiques.

Enfin, nous observons que la combinaison de classificateurs indépendants (par exemple, système de contre-mesures et VAL) rend l'évaluation peu gênant. Bien que récemment le cadre EPS [48] apparaît comme la première approche pour résoudre le problème de spoofing et les contre-évaluations, ce cadre implique nécessairement combinaison des systèmes (de fusion) au niveau de la partition en cette thèse ne considère que les configurations intégrées au niveau de décision . La méthodologie d'évaluation de Tabula Rasa se adapte pour l'évaluation du système d'VAL traditionnels (profils par exemple DET), mais nous reconnaissons que ce n'est pas une approche standardisée avec de possibles limitations/faiblesses et avec la difficulté supplémentaire pour interpréter les résultats. Il est donc nécessaire de développer des indicateurs standardisés pour évaluer contre-mesure intégrée.

B.4.4 Pensées finales

Le spoofing et les contre-mesures sont loin d'être un champ de recherche mature. Les mécanismes à l'origine le spoofing ne sont pas entièrement comprises et les améliorations constantes dans les enregistreurs portables de haute qualité font de la détection de la parole rejoué une tâche plus difficile (nous notons qu'il n'y a presque pas de littérature sur les contre-attaques pour ces). Cette thèse suggère également le potentiel de tromper les systèmes VAL avec des signaux non vocaux. En outre, alors que les technologies de synthèse de conversion de voix et la parole sont en constante évolution, la perspicacité actuelle sur le spoofing conduit à des contre-mesures de compter au-delà de la connaissance préalable de l'attaque. Par conséquent, les contre-mesures sont inefficaces contre les attaques méconnues.

Le spoofing reste donc très bien un problème ouvert, qui ne semble pas avoir une solution définitive. Plus précisément, le problème de le spoofing durera aussi longtemps que le *effort* nécessaire pour surmonter un système VAL encourageait l'usurpateur de le faire (c.-à-d. l'effort nécessaire pour tromper le système est moins coûteux que le bien de le système visant à protéger). Cette thèse souligne l'évaluation de le spoofing en termes d'effort et de prioriser l'étude des tentatives de spoofing de niveau faible dans la recherche future.

Une question plus intéressante concerne la **orientation future de cette recherche**. Indépendamment de la modalité biométrique, cette thèse en déduit que la prochaine génération de systèmes de reconnaissance biométrique sera mis enœuvre par moyenne de MCS qui reflète le problème de la reconnaissance fiable formulé comme un problème multi-classes avec détection de valeurs aberrantes. Pour la voix, à la perspicacité actuelle sur le spoofing, le premier ensemble de systèmes sera probablement complexe (ce est à dire en définissant une classe par attaque) et infractions à la sécurité, mais avec l'avance sur la recherche dans le domaine de leur complexité va diminuer et converger progressivement le problème à deux classes conventionnelle, dans le cas hypothétique où le spoofing, pouvant être statistiquement modélisé, est considéré comme une variation intra-classe. Cette question, comme d'autres plusieurs les fondamentaux, est ouvert à la recherche, mais en tenant compte du fait que le spoofing comprend attaques allant de mimique à la non-parole, les signaux audio artificiels, cette tâche est, au moins, difficile.

Bibliography

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. In *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pages 655–658. IEEE, 1988. 39

[2] A. Adler. Biometric system security. In *Handbook of biometrics*, pages 381–402. Springer, 2008. 32

[3] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12):2037–2041, 2006. 112, 114

[4] Z. Akhtar, G. Fumera, G-L Marcialis, and F. Roli. Evaluation of serial and parallel multibiometric systems under spoofing attacks. In *Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on*, pages 283–288. IEEE, 2012. 91

[5] F. Alegre, A. Amehraye, and N. Evans. A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns. In *Proc. Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, Washington DC, USA, 2013. 45, 47, 110, 115, 116, 129, 134, 153, 182

[6] F. Alegre, A. Amehraye, and N. Evans. Spoofing countermeasures to protect automatic speaker verification from voice conversion. In *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013. 104, 106, 107, 134, 153, 155, 178, 182

[7] F. Alegre, N. Evans, T. Kinnunen, Z. Wu, and J. Yamagishi. *Anti-spoofing: voice databases*. Springer, 2014. 60, 61, 63

[8] F. Alegre, A. Janichi, and N. Evans. Re-assessing the threat of replay spoofing attacks against automatic speaker verification. In *2014 International Conference of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, 2014. 58, 143

[9] F. Alegre, G. Soldi, and N. Evans. Evasion and obfuscation in automatic speaker verification. In *39th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy. 2014. 147

[10] F. Alegre, G. Soldi, N. Evans, B. Fauve, and J. Liu. Evasion and obfuscation in speaker recognition surveillance and forensics. In *2nd International Workshop on Biometrics and Forensics (IWBF 2014)*, Malta, 2014. 2, 3, 147

[11] F. Alegre, R. Vipperla, A. Amehraye, and N. Evans. A new speaker verification spoofing countermeasure based on local binary patterns. In *Proc. Interspeech*, Lyon, France, 2013. 47, 110, 113, 114, 116, 134, 182

[12] F. Alegre, R. Vipperla, and N. Evans. Spoofing countermeasures for the protection of automatic speaker recognition from attacks with artificial signals. In *Proc. 13th Interspeech*, 2012. 29, 54, 58, 63, 69, 81, 102, 103, 134, 140, 167, 174, 182, 186

[13] F. Alegre, R. Vipperla, N. Evans, and B. Fauve. On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals. Technical Report RR-12-266, EURECOM, 2012. 53, 54, 58, 63

[14] F. Alegre, R. Vipperla, N. Evans, and B. Fauve. On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals. In *Proc. 12th EUSIPCO*, 2012. 167

[15] F. Alegre, X. Zhao, N. Evans, J. Bustard, M. Nixon, A. Hadid, W. Ketchantang, S. Picard, S. Revelin, A. Riera, et al. Tabula rasa trusted biometrics under spoofing attacks. 21, 29

[16] A. Anjos and S. Marcel. Counter-measures to photo attacks in face recognition: a public database and a baseline. In *International Joint Conference on Biometrics (IJCB),*, pages 1–7. IEEE, 2011. 60

[17] B. S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *the Journal of the Acoustical Society of America*, 55(6):1304–1312, 1974. 15

[18] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1):42–54, 2000. 20

[19] R. Auckenthaler and J. S Mason. Gaussian selection applied to text-independent speaker verification. In *Proc. Speaker Odyssey*, pages 83–88, 2001. 21

[20] G. Aversano, N. Brümmer, and M. Falcone. Evalita 2009 speaker identity verification application track - organizerś report. Reggio Emilia, Italy, 2009. 22

[21] A. Benyassine, E. Shlomot, H-Y Su, D. Massaloux, C. Lamblin, and J-P Petit. Itu-t recommendation g. 729 annex b: a silence compression scheme for use with g. 729 optimized for v. 70 digital simultaneous voice and data applications. *Communications Magazine, IEEE*, 35(9):64–73, 1997. 18

[22] B. Biggio, G. Fumera, and F. Roli. Evade hard multiple classifier systems. In *Applications of Supervised and Unsupervised Ensemble Methods*, pages 15–38. Springer, 2009. 91

[23] B. Biggio, G. Fumera, and F. Roli. Multiple classifier systems under attack. In *Multiple Classifier Systems*, pages 74–83. Springer, 2010. 91

[24] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds. A tutorial on text-independent speaker verification. *EURASIP J. Appl. Signal Process.*, 2004:430–451, January 2004. 6, 15, 16, 19, 21

[25] C. M. Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006. 88

[26] C. M. Bishop, J. Lasserre, et al. Generative or discriminative? getting the best of both worlds. *Bayesian Statistics*, 8:3–23, 2007. 99

[27] A. W. Black. Clustergen: a statistical parametric synthesizer using trajectory modeling. In *Proc. Interspeech*, 2006. 37

[28] M. Blomberg, D. Elenius, and E. Zetterholm. Speaker verification scores and acoustic analysis of a professional impersonator. In *Proc. FONETIK*, 2004. 29, 58, 63

[29] J.-F. Bonastre, D. Matrouf, and C. Fredouille. Transfer function-based voice transformation for speaker recognition. In *Proc. Speaker Odyssey*, pages 1–6, 2006. 23, 166

[30] J.-F. Bonastre, D. Matrouf, and C. Fredouille. Artificial impostor voice transformation effects on false acceptance rates. In *Proc. Interspeech*, pages 2053–2056, 2007. 58, 63, 166

[31] J.-F. Bonastre, N. Scheffer, C. Fredouille, and D. Matrouf. NIST'04 speaker recognition evaluation campaign: new LIA speaker detection platform based on ALIZE toolkit. In *NIST SRE'04*, 2004. 67, 163

[32] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher,

A. Preti, G. Pouchoulin, N. Evans, B. Fauve, and J. Mason. ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition. In *Proc. Speaker Odyssey*, volume 5, page 1, 2008. 25, 67, 163

[33] J.-F. Bonastre, F. Wils, and S. Meignier. ALIZE, a free toolkit for speaker recognition. In *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 737 – 740, 18-23, 2005. 68, 163

[34] N. Brümmer and E. De Villiers. The speaker partitioning problem. In *Proc. Speaker Odyssey*, page 34, 2010. 27

[35] J. P. Campbell. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, 1997. 7, 15, 16, 34, 159

[36] J. P. Campbell and A. Higgins. Yoho speaker verification. *Linguistic Data Consortium*, 1994. 22

[37] W. M. Campbell and K. T. Assaleh. Polynomial classifier techniques for speaker verification. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 1, pages 321–324. IEEE, 1999. 19

[38] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-carrasquillo. Support vector machines for speaker and language recognition. *Computer Speech and Language*, 20:210–229, 2006. 16, 19, 27

[39] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using gmm supervectors for speaker verification. *IEEE Signal Processing Letters*, 13:308–311, 2006. 19, 27, 66, 164

[40] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff. Svm based speaker verification using a gmm supervector kernel and nap variability compensation. In *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, page I, may 2006. 27, 67, 164

[41] M. J. Carey, E. S. Parris, and J. S. Bridle. A speaker verification system using alphanets. In *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 397–400, April 1991. 26

[42] A. Cavoukian and A. Stoianov. *Biometric encryption: A positive-sum technology that achieves strongauthentication, security and privacy.* In-

formation and Privacy Commissioner, Ontario, 2007. 32

[43] M. M. Chakka, A. Anjos, S. Marcel, R. Tronci, D. Muntoni, G. Fadda, M. Pili, N. Sirena, G. Murgia, M. Ristori, et al. Competition on counter measures to 2-d facial spoofing attacks. In *Proc. IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–6, 2011. 29

[44] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. 119, 179

[45] L.-W. Chen, W. Guo, and L.-R. Dai. Speaker verification against synthetic speech. In *7th Int. Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2010. 38

[46] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *Biometrics Special Interest Group (BIOSIG), 2012 BIOSIG-Proceedings of the International Conference of the*, pages 1–7. IEEE, 2012. 87

[47] I. Chingovska, A. Anjos, and S. Marcel. Anti-spoofing in action: Joint operation with a verification system. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 98–104, June 2013. 86, 90, 91, 92

[48] I. Chingovska, A. Anjos, and S. Marcel. Biometrics evaluation under spoofing attacks, 2014. 45, 86, 135, 143, 190

[49] CSLU. CSLU: Speaker recognition version 1.1. *Linguistic Data Consortium*, 2006. 22

[50] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko. The CHAINS corpus: CHAracterizing INdividual Speakers. *Proc of SPECOM*, pages 431–435, 2006. 22

[51] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel. Can face anti-spoofing countermeasures work in a real world scenario? *6th IAPR International Conference on Biometrics (ICB)*, 2013. 85, 88, 100

[52] P. L. De Leon, V. R. Apsingekar, M. Pucher, and J. Yamagishi. Revisiting the security of speaker verification systems against imposture using synthetic speech. In *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1798 –1801, march 2010. 38, 58, 63, 97

[53] P. L. De Leon, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi. Detection of synthetic speech for the problem of imposture. In *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4844–4847, 2011. 29, 100

[54] P. L. De Leon, M. Pucher, and J. Yamagishi. Evaluation of the vulnerability of speaker verification to syntheticspeech. In *Proc. Speaker Odyssey*, 2010. 29, 100, 109, 178

[55] P. L. De Leon, M. Pucher, J. Yamagishi, and I. Hernaez, I. andSaratxaga. Evaluation of speaker verification security and detection of HMM-basedsynthetic speech. *ITASLP*, 20(8):2280–2290, 2012. 38

[56] P. L. De Leon, B. Stewart, and J. Yamagishi. Synthetic speech discrimination using pitch pattern statistics derived from image analysis. In *Proc. 13th Interspeech*, 2012. 29, 38, 100

[57] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In *Proc. Interspeech*, volume 9, pages 1559–1562, 2009. 21

[58] N. Dehak, P. Dumouchel, and P. Kenny. Modeling prosodic features with joint factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(7):2095–2103, 2007. 18

[59] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):788–798, 2011. 20, 27, 67, 164

[60] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977. 27

[61] G. Doddington, W. Liggett, A. Martin, and M. Przybockiand D. Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. Gaithersburg, MD, 1998. National Institute of Standards and Technology. 7, 34, 159

[62] G. R. Doddington. Speaker recognition-identifying people by their voices. *Proceedings of the IEEE*, 73(11):1651–1664, 1985. 15

[63] N. M. Duc and B. Q. Minh. Your face is not your password face authentication bypassing lenovo–asus–toshiba. *Black Hat Briefings*, 2009.

4, 158

[64] R. B. Dunn, T. F. Quatieri, D. A. Reynolds, and J. P. Campbell. Speaker recognition from coded speech and the effects of score normalization. In *Conference Record of the Thirty-Fifth Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1562–1567. IEEE, 2001. 21

[65] A. Eriksson and P. Wretling. How flexible is the human voice?–a case study of mimicry. *Target*, 30(43.20):29–90, 1997. 34

[66] N. Evans, T. Kinnunen, and J. Yamagishi. Spoofing and countermeasures for automatic speaker verification. In *Proc. Interspeech 2013*, Lyon, France, 2013. 45, 62, 147

[67] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P De Leon. *Speaker recognition anti-spoofing*. Springer, 2014. 34, 44, 45, 48, 61, 62

[68] M. Farrús, M. Wagner, J. Anguita, and J. Hern. How vulnerable are prosodic features to professional imitators? In *Proc. Speaker Odyssey*, 2008. 29, 34

[69] M. Faundez-Zanuy. On the vulnerability of biometric security systems. *IEEE Aerospace and Electronic Systems Magazine*, 19(6):3 – 8, june 2004. 4, 158

[70] M. Faundez-Zanuy, M. HagmÃ$\frac{1}{4}$ller, and G. Kubin. Speaker verification security improvement by means of speech watermarking. *Speech Communication*, 48(12):1608 – 1619, 2006. <ce:title>NOLISP 2005</ce:title>. 30, 60

[71] B. Fauve. *Tackling variabilities in speaker verification with a focus on short durations*. PhD thesis, Swansea University, 2009. 19, 21

[72] B. G. B. Fauve, D. Matrouf, N. Scheffer, J.-F. Bonastre, and J. S. D. Mason. State-of-the-art performance in text-independent speaker verification through open-source software. *IEEE Transactions on Audio Speech and Language processing*, 15(7):1960–1968, 2007. 21, 26, 66, 67, 163, 164

[73] J. Fiérrez-Aguilar and J. O. García. *Adapted fusion schemes for multimodal biometric authentication*. J. Fiérrez Aguilar, 2006. 93

[74] F. H. Foomany, A. Hirschfield, and M. Ingleby. Toward a dynamic framework for security evaluation of voice verification systems. In *Proc. IEEE Toronto Int. Conf. Science and Technology for Humanity (TIC-*

STH), pages 22–27, 2009. 38, 60

[75] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(2):254–272, 1981. 18

[76] S. Furui. An overview of speaker recognition technology. In *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pages 1–9, 1994. 22

[77] Javier Galbally. *Vulnerabilities and attack protection in security systems based on biometric recognition.* J. Galbally-Herrero, 2009. 17, 30, 31, 32, 33

[78] J. Galbally-Herrero, S. Carballo, J. Fiérrez-Aguilar, and J. Ortega-Garcia. Vulnerability assessment of fingerprint matching based on time analysis. In *Biometric ID Management and Multimodal Communication*, pages 285–292. Springer, 2009. 31

[79] J. Galbally-Herrero, J. Fierrez-Aguilar, J. D. Rodriguez-Gonzalez, F. Alonso-Fernandez, J. Ortega-Garcia, and M. Tapiador. On the vulnerability of fingerprint verification systems to fake fingerprints attacks. In *Proceedings 40th Annual IEEE International Carnahan Conferences Security Technology*, pages 130–136. IEEE, 2006. 4, 158

[80] Guillaume Galou. Synthetic voice forgery in the forensic context: a short tutorial. In *Forensic Speech and Audio Analysis Working Group (ENFSI-FSAAWG)*, pages 1–3, September 2011. 38

[81] D. Garcia-Romero and C. Y. Espy-Wilson. Analysis of i-vector length normalization in speaker recognitionsystems. In *International Conference on Speech Communication and Technology*, pages 249–252, 2011. 20, 67, 164

[82] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallettand N. Dahlgren, and V. Zue. Timit acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*, 1993. 22

[83] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2:291–298, 1994. 19, 27

[84] G. Giacinto and F. Roli. Intrusion detection in computer networks by multiple classifier systems. In *Proceedings in 16th International Conference on Pattern Recognition*, volume 2, pages 390–393. IEEE, 2002.

91

[85] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny. Comparison of scoring methods used in speaker recognition with jointfactor analysis. In *IEEEInternational Conference on Acoustics, Speech and Signal Processing (ICASSP), 2009*, pages 4057–4060. IEEE, 2009. 21

[86] J. Godfrey and E. Holliman. Switchboard-1 release 2. *Linguistic Data Consortium*, 1997. 22

[87] K.-S. Goh, E. Y Chang, and B. Li. Using one-class and two-class svms for multiclass image annotation. *Knowledge and Data Engineering, IEEE Transactions on*, 17(10):1333–1346, 2005. 91

[88] M. Hebert. Text-dependent speaker recognition. In Jacob Benesty, M.Mohan Sondhi, and Yiteng(Arden) Huang, editors, *Springer Handbook of Speech Processing*, pages 743–762. Springer Berlin Heidelberg, 2008. 16

[89] M. Heikkilä, M. Pietikäinen, and J. Heikkilä. A texture-based method for detecting moving objects. In *BMVC*, pages 1–10, 2004. 112

[90] K. Hempstalk and E. Frank. Discriminating against new classes: One-class versus multi-classclassification. In *AI 2008: Advances in Artificial Intelligence*, pages 325–336. Springer, 2008. 86

[91] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, 1994. 18

[92] A. K Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000. 90, 93, 99

[93] A. K. Jain, K. Nandakumar, and A. Nagar. Biometric template security. *EURASIP Journal on Advances in Signal Processing*, 2008:113, 2008. 30, 32

[94] A. K. Jain, A. Ross, and S. Pankanti. Biometrics: a tool for information security. *IEEE Transactions on Information Forensics and Security*, 1(2):125–143, 2006. 30

[95] A. K. Jain and U. Uludag. Hiding biometric data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(11):1494–1498, 2003. 32

[96] H. Jin, Q. Liu, H. Lu, and X. Tong. Face detection using improved lbp

under bayesian framework. In *2004 IEEE First Symposium on Multi-Agent Security and Survivability*, pages 306–309. IEEE, 2004. 112

[97] Q. Jin, A. R Toth, T. Schultz, and A. W. Black. Voice convergin: Speaker de-identification by voice transformation. In *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3909–3912, 2009. 147, 150

[98] P. A. Johnson, B. Tan, and S. Schuckers. Multimodal fusion vulnerability to non-zero effort (spoof) imposters. In *2010 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–5. IEEE, 2010. 45, 91

[99] A. Kain and M. W. Macon. Spectral voice conversion for text-to-speech synthesis. In *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998. 39

[100] S. S. Kajarekar, H. Bratt, E. Shriberg, and R. de Leon. A study of intentional voice modifications for evading automaticspeaker recognition. In *Proc. Speaker Odyssey*, 2006. 2, 150

[101] P. Kenny. Joint factor analysis of speaker and session variability: theoryand algorithms. Technical Report 06/08-13, CRIM, 2006. 27, 67, 84, 164

[102] P. Kenny. Bayesian speaker verification with heavy-tailed priors. In *Proc. Speaker Odyssey*, page 14, 2010. 20

[103] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1435–1447, 2007. 26

[104] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Speaker and session variability in gmm-based speaker verification. *IEEE Transactions on Audio Speech, and Language Processing*, 15(4):1448–1460, 2007. 5, 19

[105] T. Kinnunen and H. Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52(1):12 – 40, 2010. 16, 19, 20

[106] T. Kinnunen, 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 *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*

(ICASSP), pages 4401–4404, 2012. 29, 81, 84

[107] T. Kitamura. Acoustic analysis of imitated voice produced by a professional impersonator. pages 813–816, 2008. 34

[108] J. Kittler and F. Roli. *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21-23, 2000 Proceedings*, volume 1857. Springer, 2000. 91

[109] M. Kockmann, L. Ferrer, L. Burget, and J. Cernocký. ivector fusion of prosodic and cepstral features for speaker verification. In *Proc. Interspeech*, pages 265–268. ISCA, 2011. 18

[110] J. Komulainen, A. Hadid, M. Pietikäinen, A. Anjos, and S. Marcel. Complementary countermeasures for detecting scenic face spoofingattacks. *6th IAPR International Conference on Biometrics (ICB)*, 2013. 87

[111] B. Krawczyk and M. Woźniak. Combining diverse one-class classifiers. In *Hybrid Artificial Intelligent Systems*, pages 590–601. Springer, 2012. 91

[112] L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationshipwith the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003. 93

[113] H. J. Künzel, J. Gonzalez-Rodriguez, and J. Ortega-García. Effect of voice disguise on the performance of a forensic automatic speaker recognition system. In *Proc. Speaker Odyssey*, 2004. 150

[114] Y. Lau, D. Tran, and M. Wagner. Testing voice mimicry with the yoho speaker verification corpus. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 907–907. Springer, 2005. 34

[115] Y.W. Lau, M. Wagner, and D. Tran. Vulnerability of speaker verification to voice mimicking. In *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing*, pages 145–148. IEEE, 2004. 34

[116] H. Li, B. Ma, and K. A. Lee. Spoken language recognition: From fundamentals to practice. *Proceedings of the IEEE*, 101(5):1136–1159, 2013. 16

[117] Haizhou Li and Bin Ma. Techware: Speaker and spoken language recognition resources [best of the web]. *IEEE Signal Processing Magazine*,

27(6):139–142, 2010. 16

[118] P. Li, Y. Fu, U. Mohammed, J. H. Elder, and S. J. D. Prince. Probabilistic models for inference about identity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):144–157, 2012. 20, 27, 67, 164

[119] J. Lindberg, M. Blomberg, et al. Vulnerability in speaker verification-a study of technical impostor techniques. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 3, pages 1211–1214, 1999. 29, 35, 38, 58, 63

[120] Z.-H. Ling, X.-J. Xia, Y. Song, L.-H. Yang, C.-Y.and Chen, and L.-R. Dai. The USTC system for Blizzard Challenge 2012. In *Blizzard Challenge workshop*, 2012. 37

[121] H. Luo. Optimization design of cascaded classifiers. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 480–485. IEEE, 2005. 93

[122] J. Määttä, A. Hadid, and M. Pietikäinen. Face spoofing detection from single images using texture and local shape analysis. *Biometrics, IET*, 1(1):3–10, 2012. 29

[123] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar. *Handbook of fingerprint recognition*. springer, 2009. 31

[124] E. Marasco, Y. Ding, and A. Ross. Combining match scores with liveness values in a fingerprint verification system. In *2012 IEEE FifthInternational Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 418–425. IEEE, 2012. 92

[125] E. Marasco, P. Johnson, C. Sansone, and S. Schuckers. Increase the security of multibiometric systems by incorporatinga spoofing detection algorithm in the fusion mechanism. In *Multiple Classifier Systems*, pages 309–318. Springer, 2011. 92

[126] S. Marcel, C. McCool, P. M. Ahonen, et al. Mobile biometry (mobio) face and speaker verification evaluation. In *Proceedings of the 20th International Conference on Pattern Recognition*, 2010. 23

[127] J. Mariéthoz and S. Bengio. Can a professional imitator fool a GMM-based speaker verificationsystem? IDIAP Research Report (No. Idiap-RR 05-61), 2006. 34

[128] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki.

The det curve in assessment of detection task performance. In *Proc. EuroSpeech*, pages 1895–1898, 1997. 24

[129] A. Martin and C. Greenberg. The nist 2010 speaker recognition evaluation. *Proc. Interspeech*, pages 2726–2729, 2010. 22

[130] T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi. On the security of HMM-based speaker verification systems against imposture using synthetic speech. In *Proc. EUROSPEECH*, 1999. 29, 38, 58, 62, 63

[131] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai. Speech synthesis using HMMs with dynamic features. In *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1996. 38

[132] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai. Voice characteristics conversion for HMM-based speech synthesis system. In *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1997. 38

[133] D. Matrouf, J.-F. Bonastre, and J. P. Costa. Effect of impostor speech transformation on automatic speaker recognition. *Biometrics on the Internet*, page 37, 2005. xvi, 29, 69, 100, 103, 104, 105, 110, 151, 166

[134] D. Matrouf, J.-F. Bonastre, and C. Fredouille. Effect of voice transformation on impostor acceptance. In *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, 2006. 40, 41, 42, 43

[135] D. Matrouf, N. Scheffer, B. G. B. Fauve, and J.-F. Bonastre. A straightforward and efficient implementation of the factor analysis model for speaker verification. In *Proc. Interspeech*, pages 1242–1245, 2007. 19, 21, 67, 164

[136] T. Matsui and S. Furui. Likelihood normalization for speaker verification using a phoneme-and speaker-independent model. *SPCOM*, 17(1-2):109–116, Aug. 1995. 38

[137] O. Mazhelis and S. Puuronen. Combining one-class classifiers for mobile-user substitution detection. In *ICEIS (4)*, pages 130–137, 2004. 91

[138] S. Meignier, T. Merlin, C. Lévy, A. Larcher, E. Chartonand J-F. Bonastre, L. Besacier, J. Farinas, and B. Ravera. Mistral: plateforme open source d'authentification biométrique. *les actes de Journées d'Etudes*

sur la Parole (JEP), Avignon, France, pages 81–84, 2008. 25

[139] P. Mermelstein. *Distance Measures for Speech Recognition, psychological and instrumental*. Pattern recognition and artificial intelligence. Academic Press, New York, 1976. 26

[140] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, 9(5):453–467, 1990. 70, 168

[141] Z. Wu N. Evans, F. Alegre and T. Kinnunen. *Automatic speaker verification spoofing: voice conversion*. Springer, 2014. 43

[142] NIST. The nist speaker recognition evaluations, 1997-2014. 15

[143] NIST. The nist year 2005 speaker recognition evaluation plan, 2005. 23

[144] NIST. The nist year 2006 speaker recognition evaluation plan, 2006. 23

[145] A. Ogihara and A. Shiozaki. Discrimination method of synthetic speech using pitch frequency againstsynthetic speech falsification. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 88(1):280–286, 2005. 29, 38, 100

[146] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996. 111

[147] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classificationwith local binary patterns. *Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. 111, 112, 178

[148] J. Ortega-Garcia, J. Fierrez-Aguilar, F. Alonso-Fernandez, J. Galbally-Herreroand M. R. Freire, J. Gonzalez-Rodriguez, C. Garcia-Mateo, J-L Alba-Castro, E. Gonzalez-Agulla, E. Otero-Muras, et al. The multiscenario multienvironment biosecure multimodal database (bmdb). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1097–1111, 2010. 23

[149] P.-Y. Oudeyer. The production and recognition of emotions in speech: Featuresand Algorithms. *International Journal of Human-Computer Studies*, 59:157 – 183, 2003. 56

[150] D. Pearce and H. Hirsch. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy con-

ditions. *ISCA ITRW ASR2000*, pages 29–32, 2000. 22

[151] J. Pelecanos and S. Sridharan. Feature warping for robust speaker verification. pages 213–218, 2001. 18, 26

[152] B. L. Pellom and J. H. L. Hansen. An experimental study of speaker verification sensitivity to computer voice-altered imposters. In *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 837–840, 1999. 29, 39, 62

[153] P. Perrot, G. Aversano, R. Blouet, M. Charbit, and G. Chollet. Voice forgery using ALISP : Indexation in a Client Memory. In *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 17 – 20, 2005. 29, 58, 63

[154] P. Perrot, G. Aversano, and G. Chollet. Voice disguise and automatic detection: review and perspectives. In *Progress in nonlinear speech processing*, pages 101–117. Springer, 2007. 149

[155] P. Perrot, M. Morel, J. Razik, and G. Chollet. Vocal forgery in forensic sciences. In *Forensics in Telecommunications, Information and Multimedia*, pages 179–185. Springer, 2009. 2, 150

[156] M. A. Przybocki, A. F. Martin, and A. N. Le. Nist speaker recognition evaluations utilizing the mixer corpora–2004, 2005, 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):1951–1959, 2007. 23

[157] T. F. Quatieri. *Discrete-Time Speech Signal Processing Principles and Practice*. Prentice-Hall, Inc., 2002. 38

[158] N. K. Ratha, J. H. Connell, and R. M. Bolle. An analysis of minutiae matching strength. In *Proc. 3rd AVBPA*, pages 223–228, 2001. 30, 32

[159] N. K. Ratha, J. H. Connell, and R. M. Bolle. Enhancing security and privacy in biometrics-based authentication systems. *IBM Systems Journal*, 40(3):614–634, 2001. 2

[160] D. A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech communication*, 17(1):91–108, 1995. 19

[161] D. A. Reynolds. Channel robust speaker verification via feature mapping. In *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages II–53. IEEE, 2003. 18, 20, 26

[162] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3):19 – 41, 2000. 16, 19, 26, 27, 104

[163] G. Ritter and M. T. Gallegos. Outliers in statistical pattern recognition and an application to automatic chromosome classification. *Pattern Recognition Letters*, 18(6):525–539, 1997. 88

[164] R. Rodman. Speaker recognition of disguised voices: a program for research. In *Proceedings of the Consortium on Speech Technology in Conjunction with the Conference on Speaker Recognition by Man and Machine: Directions for Forensic Applications, Ankara, Turkey, COST250 Publishing Arm*, pages 9–22. Citeseer, 1998. 149

[165] R. N. Rodrigues, N. Kamat, and V. Govindaraju. Evaluation of biometric spoofing in a multimodal system. In *2010 Fourth IEEEInternational Conference on Biometrics: Theory Applications and Systems (BTAS)*, pages 1–5. IEEE, 2010. 91

[166] R. N. Rodrigues, L. L. Ling, and V. Govindaraju. Robustness of multimodal biometric fusion methods against spoof attacks. *Journal of Visual Languages & Computing*, 20(3):169–179, 2009. 91

[167] F. Roli. Mini tutorial: Three hours on multiple classifier systems. Lecture Notes - School on the Analysis of Patterns, 2009. 92, 95

[168] F. Roli, G. Giacinto, and G. Vernazza. Methods for designing multiple classifier systems. In *Multiple Classifier Systems*, pages 78–87. Springer, 2001. 95

[169] A. Ross, K. Nandakumar, and A. K. Jain. Introduction to multibiometrics. In *Handbook of biometrics*, pages 271–292. Springer, 2008. 91

[170] A. Roy, M. Magimai-Doss, and S. Marcel. A fast parts-based approach to speaker verification using boosted slice classifiers. *IEEE Transactions on Information Forensics and Security*, 7(1):241–254, 2012. 111

[171] M. Russell and R. Moore. Explicit modelling of state occupancy in hidden markov models for automatic speech recognition. In *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5–8, 1985. 70, 168

[172] M. Sahidullah and G. Saha. Comparison of speech activity detection techniques for speaker recognition. *arXiv e-preprint*, Oct. 2012. 18, 149

[173] T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda. A robust speaker verification system against imposture using an HMM-based speech synthesis system. In *In proc. EUROSPEECH*, 2001. 38, 100

[174] M. Schmidt and H. Gish. Speaker identification via support vector classifiers. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings.*, volume 1, pages 105–108. IEEE, 1996. 19

[175] W. Shang and M. Stevenson. Score normalization in playback attack detection. In *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 1678–1681. IEEE, 2010. 36

[176] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke. Modeling prosodic feature sequences for speaker recognition. *Speech Communication*, 46(3):455–472, 2005. 18

[177] S. Siddiq, T. Kinnunen, M. Vainio, and S. Werner. Intonational speaker verification: A study on parameters and performanceunder noisy conditions. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4777–4780. IEEE, 2012. 18

[178] A. Solomonoff, W.M. Campbell, and I. Boardman. Advances in channel compensation for svm speaker recognition. In *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 629 – 632, 18-23, 2005. 19

[179] P. L. Sordo Martinez, B. Fauve, A. Larcher, and J. S. D. Mason. Speaker verification performance with constrained durations. In *2014 International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, March 2014. 135, 184

[180] C. Soutar, R. Gilroy, and A. Stoianov. Biometric system performance and security. In *Conf. IEEE Auto. Identification Advanced Technol*, 1999. 31

[181] Y. Stylianou, O. Cappé, and E. Moulines. Continuous probabilistic transform for voice conversion. *ITSAP*, 6(2):131–142, 1998. 39

[182] V. Takala, T. Ahonen, and M. Pietikäinen. Block-based methods for image retrieval using local binary patterns. In *Image Analysis*, pages 882–891. Springer, 2005. 112

[183] D. Talkin. A robust algorithm for pitch tracking (rapt). *Speech coding and synthesis*, 495:518, 1995. 118, 178

[184] D. M. J. Tax. One-class classification. 2001. 101

[185] D. M. J. Tax and R. P. W. Duin. Growing a multi-class classifier with a reject option. *Pattern Recognition Letters*, 29(10):1565–1570, 2008. 88

[186] R. Togneri and D. Pullella. An overview of speaker identification: accuracy and robustness issues. *IEEE Circuits and Systems Magazine*, 11(2):23–61, 2011. 16

[187] T. Tomoki and K. Tokuda. A speech parameter generation algorithm considering global variance for hmm-based speech synthesis. *IEICE TRANSACTIONS on Information and Systems*, 90(5):816–824, 2007. 100

[188] M. Turtinen, M. Pietikainen, and O. SilvÉn. Visual characterization of paper using isomap and local binary patterns. *IEICE transactions on information and systems*, 89(7):2076–2083, 2006. 112

[189] P. Tuyls, A. H. M. Akkermans, T. A. M. Kevenaar, G.-J. Schrijen, A. M. Bazen, and R. N. J. Veldhuis. Practical biometric authentication with template protection. In *Audio-and Video-Based Biometric Person Authentication*, pages 436–446. Springer, 2005. 32

[190] J. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995. 27

[191] V. N. Vapnik and V. Vapnik. *Statistical learning theory*, volume 2. Wiley New York, 1998. 19

[192] J. Villalba and E. Lleida. Speaker verification performance degradation against spoofing and tampering attacks. In *FALA workshop*, pages 131–134, 2010. 4, 29, 36, 38, 58, 63, 150, 158

[193] J. Villalba and E. Lleida. Detecting replay attacks from far-field recordings on speaker verification systems. In *Biometrics and ID Management*, pages 274–285. Springer, 2011. 36

[194] R. Vogt and S. Sridharan. Experiments in session variability modelling for speaker verification. In *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages I–I. IEEE, 2006. 19

[195] Z.-F. Wang, G. Wei, and Q.-H. He. Channel pattern noise based playback attack detection algorithm for speaker recognition. In *Proc. IEEE Int. Conf. Machine Learning and Cybernetics (ICMLC)*, 2011. 36

[196] H. Wu, Y. Wang, and J. Huang. Blind detection of electronic disguised voice. In *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3013–3017, 2013. 150

[197] R.-S. Wu and W.-H. Chung. Ensemble one-class support vector machines for content-based image retrieval. *Expert Systems with Applications*, 36(3):4451–4459, 2009. 91

[198] Z. Wu, E.S. Chng, and H. Li. Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In *Proc. 13th Interspeech*, 2012. 29, 38, 43, 47, 100, 153

[199] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li. An overview of spoofing and countermeasures for automatic speaker verification. *ELSEVIER Speech Communication Journal*, 2014. 32, 34, 44, 48

[200] Z. Wu, T. Kinnunen, E.S. Chng, H. Li, and E. Ambikairajah. A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, pages 1–5. IEEE, 2012. 29, 42, 43, 100, 109, 178

[201] Z. Wu and H. Li. Voice conversion and spoofing attack on speaker verification systems. In *APSIPA*, 2013. 40

[202] L. Xu, A. Krzyzak, and C. Y. Suen. Methods of combining multiple classifiers and their applicationsto handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3):418–435, 1992. 93

[203] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai. Analysis of Speaker Adaptation Algorithms for HMM-based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm. *IEEE transactions on Audio, Speech & Language Processing*, 17(1):66–83, 2009. 70, 168

[204] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals. Robust speaker adaptive HMM based Text-to-Speech Synthesis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 17(6):1208–1230, 2009. 70, 168

[205] P. Yang, Y. Hwa Yang, B. B Zhou, and A. Y Zomaya. A review of ensemble methods in bioinformatics. *Current Bioinformatics*, 5(4):296–308, 2010. 91

[206] M. M. Yeung and S. Pankanti. Verification watermarks on fingerprint recognition and retrieval. *Journal of Electronic Imaging*, 9(4):468–476, 2000. 32

[207] S. Yoon, J. Feng, and A.K. Jain. Altered fingerprints: Analysis and detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3):451–464, 2012. 148, 149

[208] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Simultaneous modeling of spectrum, pitch and duration in HMM-basedspeech synthesis. In *In proc. EUROSPEECH*, 1999. 37

[209] H. Yu. Svmc: Single-class classification with support vector machines. In *IJCAI*, pages 567–574. Citeseer, 2003. 101

[210] H. Zen, K. Tokuda, and A. W. Black. Statistical parametric speech synthesis. *SPCOM*, 51(11):1039–1064, 2009. 37

[211] C. Zhang and T. Tan. Voice disguise and automatic speaker recognition. *Forensic science international*, 175(2):118–122, 2008. 150

[212] G. Zhang, X. Huang, S.Z. Li, Y. Wang, and X. Wu. Boosting local binary pattern (lbp)-based face recognition. In *Advances in biometric person authentication*, pages 179–186. Springer, 2005. 112

[213] C. Zhi and Z. Ling-hua. Voice conversion based on Genetic Algorithms. In *12th IEEE International Conference on Communication Technology (ICCT)*, pages 1407 –1409, nov. 2010. 56

[214] Z.-H. Zhou, F. Roli, and J. Kittler. *Multiple Classifier Systems: 11th International Workshop, MCS 2013 Nanjing, China, May 15-17, 2013 Proceedings*, volume 7872. Springer, 2013. 91

[215] G. Zuo and W. Liu. Genetic algorithm based RBF neural network for voice conversion. In *Fifth World Congress on Intelligent Control and Automation*, volume 5, pages 4215 – 4218, june 2004. 56