Towards robust and privacy-preserving speaker verification systems

Aymen Mtibaa

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2022IPPAS002

Thèse de doctorat

TOWARDS ROBUST AND PRIVACY-PRESERVING SPEAKER VERIFICATION SYSTEMS

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom SudParis

École doctorale n°626: École doctorale de l'Institut Polytechnique de Paris (IPP)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Paris, le 14 Février 2022, par

MTIBAA AYMEN

Composition du Jury :

Prof. Christophe Rosenberger Ecole Nationale Supérieure d'Ingénieurs de Caen- GREYC	Président, Rapporteur
Prof. Jan Cernocky Faculty of Information Technology, Brno University of Technology (FIT BUT)	Rapporteur
Prof. Isabel Trancoso INESC-ID / IST, University Lisbon	Examinateur
Dr. Sanjay Kanade TSSM's, Bhivarabai Sawant College of Engineering Research	Examinateur
Prof. Jérôme Boudy Télécom SudParis	Directeur de thèse
Maître de conférences. Dijana Petrovska-Delacrétaz Télécom SudParis	Encadrant de thèse

Abstract

Speaker verification systems are a key technology in many devices and services like smartphones, intelligent digital assistants, and banking applications. Moreover, during the COVID-19 pandemic, access control systems based on fingerprint scanners or keypads increase the risk of virus propagation. Therefore, companies are now rethinking their employee access control systems and considering touchless authorization technologies, such as speaker verification systems.

However, speaker verification systems require that the access system stores the speakers' models and has access to the recordings or features derived from the speakers' voices during the authentication. This process raises some concerns regarding the privacy of the user and the protection of such sensitive biometric data. An adversary can steal speakers' models, features, or recordings from the access system and use this biometric information to impersonate the genuine user and gain unauthorized access. Moreover, when dealing with speech data, we are in front of additional privacy concerns. In case the speech data are stolen, several personal information related to the speaker's identity, gender, age, or health status could be extracted. Therefore, speaker verification systems should be improved in a way that preserves speaker privacy and ensures the protection of biometric information stored (i.e.biometric reference) or provided during the authentication.

In this context, the present PhD Thesis addresses the privacy and security issues for speaker verification systems based on Gaussian mixture models (GMM), i-vector, and x-vector as speaker modeling. The objective is the development of speaker verification systems that perform biometric verification while preserving the privacy and the security of the user. To that end, we have proposed biometric protection schemes for speaker verification systems to achieve the privacy requirements (revocability, unlinkability, and irreversibility) described in the stan-

dard ISO/IEC IS 24745 on biometric information protection and to improve the robustness of the systems against different attack scenarios.

In this thesis, we first presented the existing biometric information protection schemes that address the privacy-preserving for speaker recognition systems. We classified the schemes into three categories (i) cryptography-based schemes; (ii) cancelable based schemes; and (iii) hybrid based schemes. For this thesis, we are focusing on cancelable biometrics where intentional and systematically repeatable distortion is applied to biometric features or references in order to protect sensitive user data.

In order to improve the privacy and security for speaker verification systems based on GMM and i-vectors, we proposed a cancelable privacy-preserving based on two steps: (i) the extraction of a binary representation of the speaker derived from his/her biometric reference and (ii) the protection of the binary representation using a shuffling scheme that randomizes the binary representation with the help of a shuffling key.

The transformation of the speaker's binary representation with the shuffling scheme makes it possible to generate from the same biometric sample different versions of protected biometric references (revocability) that cannot be linked to the user (Unlinkability). These properties ensure the privacy of the user when he is enrolled in different applications using the same biometric sample (prevents cross-matching), and in case the user's protected biometric reference is compromised, it will be revoked and renewed. Furthermore, the cancelable scheme makes it possible to simultaneously achieve the privacy requirements while maintaining the biometric verification performance. Due to the shuffling scheme, the biometric performance of the privacy-preserving speaker verification systems outperforms that of the baseline (unprotected) systems. Regarding security, the cancelable scheme makes the systems robust against different attack scenarios. As an example, in case the user's biometric data are compromised, the systems are robust with a false acceptance rate equal to zero. However, in case the user's shuffling key is compromised, a degradation in terms of false acceptance rate was observed. This degradation is related to the loss of biometric performance when transforming the speaker's model into binary representation before applying the shuffling scheme.

Therefore, the cancelable privacy-preserving scheme was improved by propos-

ing a binarization approach of the speaker biometric reference based on deep neural nets autoencoder. This approach transforms the speaker's biometric reference into binary representation while maintaining the biometric performance and makes it possible to control the dimension of the binary representation. In addition, we have proposed to apply secure sketch error correction code (EEC) to the binary representation protected with the shuffling scheme. The goal was to take advantage of the shuffling transformation and error correction to improve the security and the biometric performance.

The improved cancelable scheme was used to develop a privacy-preserving speaker verification system based on x-vectors extracted from a Time Delay Neural Network (TDNN). Protection of the x-vector is performed by first transforming it into binary representations using the binarization approach based on the autoencoder on top of the TDNN. Then, cancelable x-vector is generated by transforming the binary x-vector with the shuffling scheme. This transformation allows achieving privacy requirements. Next, secure sketch error correction is applied to the cancelable x-vector in order to manage the biometric variability, which allows improving the security and the biometric performance of the system.

The protection of x-vectors with the described cancelable scheme allows the processing of speaker verification in a protected domain without revealing personal information about the user. The speaker verification system based on cancelable x-vectors achieves the privacy requirements and outperforms the biometric performance of the unprotected x-vector system. An EER=0.1% was reported compared to EER=3.12% for the baseline x-vectors. Moreover, the system is robust against stolen biometric, stolen token, and brute force attacks with a FAR=0. In addition, due to the binarization approach that maintains the biometric performance and the combination of shuffling with the error correction code, the system is robust to the stolen shuffling key scenario. For the unprotected x-vectors system, the biometric performance in terms of EER=3.12% (FAR=FRR=3.12%). For the proposed privacy-preserving x-vector system based on four enrollment utterances, at FRR=3.12%, the FAR=0 in the legitimate scenario and the FAR=1.94% for the stolen shuffling key scenario. For the privacy-preserving system based on one enrollment utterance, the system outperforms the baseline system in the legitimate scenario with EER=0.1%. However, a slight degradation in terms of FAR

was observed for the stolen key scenario. A FAR=4.1% was reported compared to 3.12% for the baseline system.

Compared to the majority of research on voice biometric protection based on cancelable schemes, the proposed privacy-preserving scheme makes it possible to simultaneously achieve privacy requirements and maintains the performance of the unprotected system in legitimate and stolen key scenarios.

Finally, during this thesis, we evaluate the proposed privacy-preserving biometric systems on common and standardized assessments using public databases to contribute reproducible research. The evaluation of privacy-preserving systems starts by reporting the biometric performance of the unprotected systems for a fair comparison with the performance of the proposed protected systems. Then, the privacy is evaluated according to the requirements described on the ISO/IEC 24745 for biometric information protection. Besides, the security of the systems is evaluated against different attack scenarios dedicated for biometric systems based on biometric protection schemes.

Keywords:

Speaker Verification system, Privacy-Preserving, Security, Biometric Information Protection.

Résumé

Les systèmes de vérification du locuteur sont une technologie clé dans de nombreux appareils et services tels que les smartphones, les assistants numériques intelligents et les applications bancaires. Pendant la pandémie de COVID-19, les systèmes de contrôle d'accès basés sur des lecteurs d'empreintes digitales ou des claviers augmentent le risque de propagation du virus. Par conséquent, les entreprises repensent maintenant leurs systèmes de contrôle d'accès des employés et envisagent des technologies d'autorisation sans contact, telles que les systèmes de vérification des locuteurs.

Cependant, les systèmes de vérification du locuteur exigent que le système d'accès sauvegarde les modèles des locuteurs et ait accès aux enregistrements ou aux caractéristiques dérivées des voix des locuteurs lors de l'authentification. Ce processus soulève certaines préoccupations concernant le respect de la vie privée de l'utilisateur et la protection de ces données biométriques sensibles. Un adversaire peut voler les informations biométriques et imiter l'identité de vrai utilisateur pour obtenir un accès non autorisé. De plus, lorsqu'il s'agit de données vocales, nous sommes confrontés à des problèmes supplémentaires de confidentialité et de respect de vie privée. À partir des enregistrements vocaux, plusieurs informations personnelles liées à l'identité, au sexe, à l'âge ou à l'état de santé du locuteur peuvent être extraites. Par conséquent, les systèmes de vérification du locuteur devraient être améliorés de manière à respecter la vie privée du locuteur et à assurer la protection des informations biométriques stockées ou fournies lors de l'authentification.

Dans ce contexte, la présente thèse de doctorat aborde les problèmes de protection des données biométriques, le respect de vie privée et la sécurité pour les systèmes de vérification du locuteur basés sur les modèles de mélange gaussien

(GMM), i-vecteur et x-vecteur comme modélisation du locuteur. L'objectif est le développement de systèmes de vérification du locuteur qui effectuent une vérification biométrique tout en respectant la vie privée et la protection des données biométriques de l'utilisateur. Pour cela, nous avons proposé des schémas de protection biométrique afin de répondre aux exigences de protection des données biométriques (révocabilité, diversité, et irréversibilité) décrites dans la norme ISO/IEC IS 24745 et pour améliorer la robustesse des systèmes contre différents scénarios d'attaques.

Dans cette thèse, nous avons d'abord présenté les schémas de protection des informations biométriques existants pour les systèmes de reconnaissance du locuteur. Nous avons classé les schémas en trois catégories, (i) les schémas basés sur la cryptographie ; (ii) les schémas basés sur les transformations révocables ; et (iii) les schémas hybrides . Pour cette thèse, nous nous concentrons sur la biométrie révocable où une transformation intentionnelle et systématiquement reproductible est appliquée sur les caractéristiques ou les références biométriques afin de protéger les données sensibles des utilisateurs.

Pour les systèmes de vérification du locuteur basés sur le GMM et i-vecteur, nous avons proposé un schéma révocable de protection des données biométriques basé sur deux étapes : (i) extraction d'une représentation binaire du locuteur dérivée de sa référence biométrique et (ii) protection de la représentation binaire en utilisant un schéma de permutation qui randomise la représentation binaire en utilisant une clé spécifique pour chaque utilisateur.

La transformation de la représentation binaire du locuteur avec le schéma de permutation permet de générer à partir d'un même échantillon biométrique différentes versions de références biométriques protégées et révocables qui ne peuvent pas être liées à l'utilisateur. Ces propriétés garantissent la protection de la vie privée de l'utilisateur lorsqu'il est inscrit dans différentes applications en utilisant le même échantillon biométrique. Aussi en cas ou la référence biométrique protégée de l'utilisateur est compromise, elle sera possible de la remplacer. En outre, le schéma proposé permet de répondre simultanément aux exigences de protection de vie privée tout en maintenant les performances de vérification biométrique. Aussi, l'évaluation de la sécurité montre que le schéma de protection des données biométriques proposé rend les systèmes robustes contre différents scénarios d'attaque. Par exemple, si les données biométriques de l'utilisateur sont compromises, les

systèmes sont robustes avec un taux de fausse acceptation égal à 0. Cependant, au cas où la clé de permutation de l'utilisateur serait compromise, une dégradation en termes de taux de fausse acceptation a été observée. Cette dégradation est liée à la perte de performances biométriques lors de la transformation du modèle du locuteur en représentation binaire avant d'appliquer le schéma de permutation.

Pour améliorer le schéma de protection, nous avons proposé une approche de binarisation de la référence biométrique du locuteur basée sur le réseau de neurones auto-encodeur. Cette approche transforme la référence biométrique du locuteur en une représentation binaire sans perte de performances biométriques et donne la possibilité de contrôler la dimension de la représentation binaire. De plus, nous avons proposé d'appliquer un code de correction d'erreur à la représentation binaire protégée par le schéma de permutation afin d'améliorer la sécurité et la performance biométrique.

Le schéma de protection amélioré a été appliqué pour la protection du système de vérification de locuteur basé sur les x-vecteurs extraits de réseaux à décalage temporel (Time Delay Neural Network TDNN). La protection du x-vecteur est réalisée en le transformant en représentation binaire à l'aide de l'approche basée sur l'auto-encodeur. Ensuite, le x-vecteur binaire est protégé à l'aide du schéma de permutation et passé à travers le code de correction d'erreur afin de gérer la variabilité biométrique, ce qui permet d'améliorer la sécurité du système.

La protection des x-vecteurs permet d'effectuer la vérification du locuteur dans un domaine protégé sans révéler les informations personnelles. L'évaluation par rapport aux exigences de protection des informations biométriques montre que le système de vérification du locuteur basé sue les x-vecteurs protégés répond aux exigences de protection de la vie privée. De plus, en utilisant les x-vecteurs protégés, la performance biométrique est améliorée en terme de taux d'égal erreur (EER) et passe de 3 .12% (avec x-vecteurs non-protégés) à 0.1%. En outre, grâce à l'approche de binarisation qui maintient les performances biométriques et la combinaison du schéma permutation avec le code de correction d'erreur, le système est robuste en cas où la clé de permutation est volé. Pour le système de x-vecteurs de base (non-protégé), la performance biométrique en termes de taux d'égal erreur est 3.12%. Pour le système basé sur les x-vecteurs protégés, pour un taux de faux rejet (FRR) égale à 3.12%, le taux de fausse acceptation (FAR) est 0 pour le

scénario classique et FAR=1.94% pour le scénario de clé de permutation volé.

Pour finir, au cours de cette thèse, nous avons aussi fait un pas en avant vers l'évaluation des systèmes biométriques préservant la vie privée sur des évaluations communes et standardisées utilisant des bases de données publiques pour contribuer à une recherche reproductible. L'évaluation des systèmes de protection de la vie privée commence par rapporter les performances biométriques des systèmes non protégés pour une comparaison équitable avec les performances des systèmes protégés. Ensuite, la protection de vie privée est évaluée selon les exigences décrites dans la norme ISO/IEC 24745 pour la protection des informations biométriques. En outre, la sécurité des systèmes est évaluée par rapport à différents scénarios d'attaque dédiés aux systèmes biométriques basés sur des schémas de protection biométrique.

Mots clés:
Système de Vérification de Locuteur, La Protection de la Vie Privée, Sécurité, Protection des Informations Biométriques.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my thesis advisor Dr. Dijana Petrovska, who has continuously motivated me with her wise and intelligent advice, from the moment I joined Télécom SudParis as a student engineer. She supports me with a continuous stream of intellectual, moral, and financial assistance. Her guidance helped me throughout the research and writing of this thesis. I will never forget the skype or zoom meetings full of support and advice, the remarks, and the encouragement. I am grateful for giving me the opportunity to participate in the European H2020 project and to gain experience. THANKS

I thank my thesis director, Prof. Jérôme Boudy. It was an honor to work with him. His continued supports, guidance, and insightful comments were instrumental for the success of this work.

I would also like to thank my thesis co-director Prof. Ahmed Ben Hamida, who gave me the opportunity to get in touch with the world of research and to join the Laboratory of Advanced Technologies For Medicine and Signals (ATMS). I really appreciate the confidence he has always shown in me.

I am grateful for Prof. Christophe Rosenberger and Prof. Jan "Honza" Cernocky for accepting my request to be a reviewer for this thesis. I also thank Prof. Isabel Trancoso, and Dr. Sanjay Kanade for honoring me by being a part of the jury.

I also thank my friend Mohamed Amine Hmani for his help during various stages of this thesis. His comments, suggestions and codes were quite helpful.

And as always, the best is for last; thanks to my family, for always being there whenever I need help and support . For not ceasing to encourage me in the hardest moments, and to celebrate each success. Without my parents, nothing would have been possible. Abdelwaheb, now you can call me doctor. This is for you. Thanks.

Aymen Mtibaa
Paris, September 2021

Contents

List of Figures

List of Tables

Acronyms

AD	Auxiliary Data
ASR	Automatic Speech Recognition
DNN	Deep Neural Network
ECC	Error Correction Codes
EER	Equal Error Rate
EM	Expectation-Minimization
FAR	False Acceptance Rate
FBANK	Filter Bank
FHE	Fully Homomorphic Encryption
FRR	False rejection Rate
GAN	Generative Adversarial Network
GC	Garbled Circuit
GDPR	General Data Protection Regulation
GMM	Gaussian Mixture Model
HE	Homomorphic Encryption
JFA	Joint Factor Analysis
LDA	Linear Discriminant Analysis
LPCC	Linear Predictive Cepstral Coefficients
LSH	Locality Sensitive Hashing
MAP	Maximum a Posteriori
MDS	Maximum Distance Separable
MFCC	Mel-Frequency Cepstral Coefficients
PHE	Partially Homomorphic Encryption

PI	Pseudonymous Identifier
PIC	Pseudonymous Identifier Comparator
PIE	Pseudonymous Identifier Encoder
PIR	Pseudonymous Identifier Recorder
PLDA	Probabilistic Linear Discriminant Analysis
PLP	Perceptual Linear Prediction
RBOMP	Random Binary Orthogonal Matrices Projection
RBR	Revocable Biometric Reference
SBE	Secure Binary Embeddings
SHE	Somewhat Homomorphic Encryption
SMPC	Secure Multi-Party Computation
STPC	Secure Two-Party Computation
SV	Speaker Verification
SVM	Support Vector Machine
TDNN	Time Delay Neural Network
TTS	Text-To-Speech
UBM	Universal Background Model
VAD	Voice Activity Detection
VC	Voice Conversation
WCCN	Within-Class Covariance Normalization

Chapter 1

Introduction

Contents

Maintaining security and privacy is a priority for people who safeguard their personal information. Traditional authentication methods such as passwords and PINs are no longer reliable and efficient since the user needs to remember multiple passwords and maintain multiple authentication tokens. Biometric systems have been established as new technology to mitigate the limitations and weaknesses of these traditional access methods. Biometric systems enable the authentication of individuals based on physiological characteristics "who you are" (iris, face, fingerprint) or based on behavioral characteristics "what you produce" (voice, signature) which cannot be forgotten or lost [Jain et al., 2004].

In advanced devices such as laptops, smartphones, and smartwatches, microphones are the most commonly found sensor. This allows biometric systems based on voice modality to gain prominence in the market and to be deployed more widely. Among voice-based biometric systems, speaker verification systems are increasingly ubiquitous and have become a popular technology for authenticating individuals and controlling access to different applications. Authentication of the user based on his/her voice is more convenient than entering passwords. It consists of automatically verifying who is speaking using the voice characteristics captured by a recording device.

The process of speaker verification system consists of two phases as shown in Figure 1.1. During the enrollment, the system collects voice samples from the speaker to create the enrollment biometric reference B_r. The model is then stored in a centralized database. During the verification, the probe biometric reference B_p extracted from the probe biometric sample is compared to the model associated with the claimed identity, generating a score. This score is compared to a predefined verification threshold to determine if the probe voice sample corresponds to a client or impostor user.

Such verification systems provide greater security and convenience than traditional methods of authentication. However, they are not designed to preserve the privacy of the speaker. The speaker's biometric data are transmitted and stored without protection in external databases and servers that may be compromised.

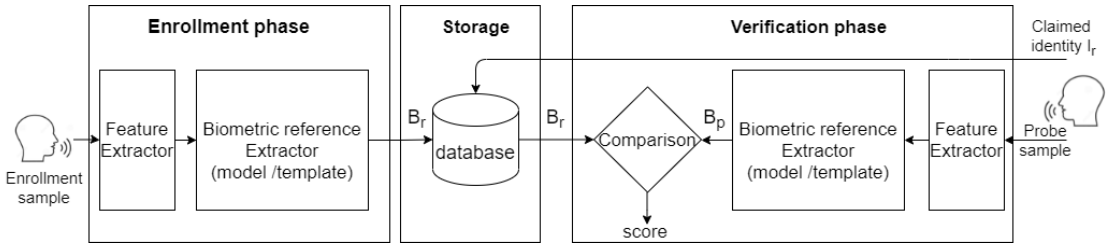


Figure 1.1: Architecture of the classical biometric system.

1.1 Privacy Issues Related to Speaker Verification Systems

Usually, we consider the unauthorized recording of our conversation through eavesdropping as a critical issue for our privacy. The current process of speaker verification system requires that the access system stores the speakers' models and has access to the recordings or features derived from the speakers' voices during the authentication. This process poses several threats to privacy and security. Speakers' models, features, or recordings can be stolen from the access system by an adversary who can use this information to create genuine recordings and gain unauthorized access. Moreover, in the case of speaker verification systems, we are in front of additional privacy concerns. Using stolen speech data, several personal information related to the speaker's identity, gender [Harb and Chen, 2005], age [Gómez García et al., 2015] or health status [Jeancolas et al., 2019] could be extracted.

In addition, unlike authentication systems based on passwords, biometric characteristics are not renewable or revocable. When using voice characteristics to authenticate, in case the target model is stolen, it becomes useless because it cannot be replaced. In a text-independent speaker verification system, where no prior constraints are considered for the spoken sentences by the speaker, once a non-target user succeeds to pre-record or synthesize the voice of the target speaker, the target voice sample is rendered useless in terms of security because the new speaker model generated from this voice sample will be the same as the compromised. For the text-dependent system, where a predefined pass-phrase is employed for verification, one possible solution is to replace the passphrase. However, in some

services and applications based on speaker verification, we are confronted with a limited choice of passphrases. For example with Google assistant, we have the choice between only *ok Google* or *hey Google*.

Another privacy issue for speaker verification systems is the cross-matching of biometric models. With the frequent use of biometrics as a form of authentication in many applications, the user could be enrolled using the same biometric instance in different access systems. For speaker verification systems, the speaker uses his/her voice to generate the enrollment models stored in different databases. Since the models are extracted from the same biometric instance, an adversary who gets access to these models could do some profiling or tracking and knows if these models correspond to the same user or not. Therefore, cross-matching between models should be prevented. In addition, if the models reveal information about the original biometric features, the adversary will be able to reconstruct synthetic features close to the original ones. As a consequence, we must ensure the irreversibility of the models.

In order to address these privacy issues, the EU General Data Protection Regulation (GDPR or Regulation 2016/679) [European Parliament and Council, 2016] classified biometric data as personal data:

> *"biometric data means personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data"* [1].

Moreover, the GDPR considered biometric data as sensitive data:

> *"Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data,* ***biometric data*** *for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited".* [2]

[1] Article 4,(14) GDPR
[2] Article 9 GDPR

> "*Personal data which are, by their nature, particularly sensitive in relation to fundamental rights and freedoms merit specific protection as the context of their processing could create significant risks to the fundamental rights and freedoms.* [3]"

These definitions mean that the processing of biometric data should take into consideration the right of privacy preservation. As a result, the traditional speaker verification system should be enhanced by new approaches that ensure the protection of sensitive personal data stored in the databases or provided during the verification process to guarantee the user's privacy.

Therefore, this thesis addresses the security and privacy issues for speaker verification systems. The objective is the development of speaker verification systems that perform biometric verification while preserving the privacy and the security of the user. More specifically, we will propose biometric protection schemes that improve the security and privacy of the systems and allow the processing of speaker verification in a protected domain without revealing personal information about the user.

1.2 Biometric Information Protection

The privacy concerns related to traditional biometric systems presented in the previous section have led to the development of the standard ISO/IEC IS 24745 on biometric information protection [ISO/IEC JTC1 SC27 Security Techniques, 2011]. This standard provides guidance for the protection of biometric information and presents an architecture of biometric systems based on the protection of biometric information. This architecture permits the generation of revocable biometric references (RBR). Revocability involves the generation of unlinkable biometric references from the same biometric characteristics.

An overview of the architectural aspects of biometric systems based on biometric information protection is presented in Figure 1.2. During the enrollment phase, a module known as pseudonymous identifier encoder (PIE) takes the enrollment features as an input and generates the revocable or renewable biometric

[3] Recital 51 GDPR

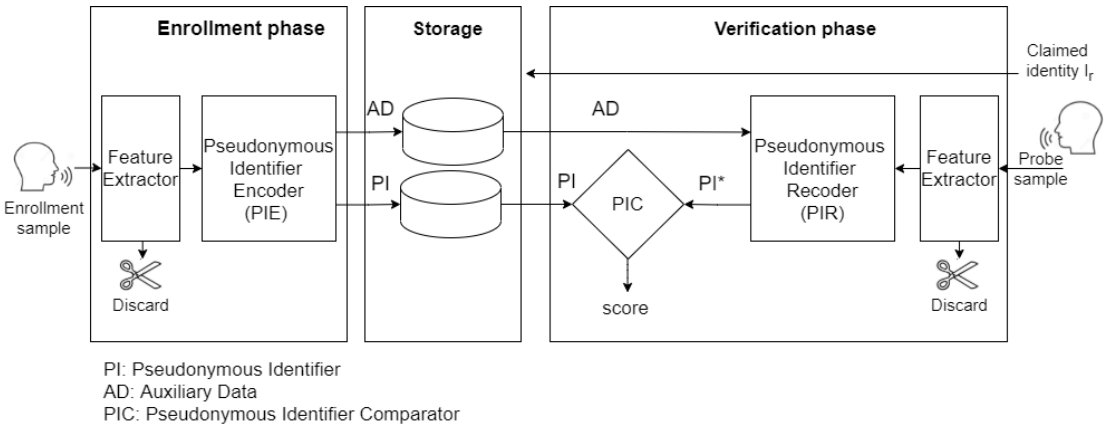


Figure 1.2: Architecture of biometric system based on the biometric information protection [ISO/IEC JTC1 SC27 Security Techniques, 2011].

reference (RBR) consisting of two elements: the pseudonymous identifier (PI) and the auxiliary data (AD). PI represents the protected biometric reference. It is the equivalent of B_r on classical biometric architecture but in a protected domain. AD is subject-dependent data that is part of the revocable biometric reference and may be required to generate the pseudonymous identifier during the verification. When the RBR is generated, the captured biometric sample and the extracted features can be securely disposed of. Then the RBR is stored, the PI and AD may be separated and stored in different databases.

During the verification phase, a module called pseudonymous identifier recorder (PIR) takes as input the probe biometric features and the stored AD to generate a protected probe biometric reference (PI*). Subsequently, a pseudonymous identifier comparator (PIC) compares the protected biometric reference PI generated during the enrollment and the probe protected biometric reference PI* and returns a similarity score. This score is then compared to a verification threshold.

Based on this new architecture, unprotected (original) biometric references are neither stored in the database nor provided in raw during the verification process. Using the modules PIE and PIR, revocable and protected biometric references are extracted and the biometric comparison is performed on the protected domain without revealing biometric information about the user.

According to the standard ISO/IEC 24745 [ISO/IEC JTC1 SC27 Security Techniques, 2011] for biometric information protection, biometric system is con-

sidering privacy-preserving when the following requirements are achieved:

Revocability: from the same biometric sample, it must be possible to generate different versions of protected biometric references. In case the subject's protected biometric reference is compromised, it will be revoked and renewed.

Unlinkability: given the same biometric sample, it must be feasible to generate different protected biometric references in a way that they cannot be linked to each other or to the subject from which they were derived.

Non-invertibility: original biometric cannot be recovered if the protected biometric reference is compromised.

Biometric performance: the protection of biometric reference should not degrade the biometric performance compared to the unprotected system.

1.3 Motivation and Objectives of The Thesis

The research carried out in this thesis has been mainly motivated by the following observations from the state-of-the-art.

Motivation 1: The development of privacy-preserving biometric systems according to the requirements established by the ISO/IEC IS 24745 standard on biometric information protection is currently a research challenge. In this direction, different biometric protection schemes have been designed for biometric systems based on face, iris, and fingerprint modalities [Kumar et al., 2020], [Rathgeb and Uhl, 2011]. However, there is a lack of protection schemes that can be adapted and applied for speaker verification systems where the biometric reference is represented with a model rather than a template. As a consequence, the first objective of this thesis was:

Objective 1: Developing methodologies to protect speaker verification systems in order to achieve privacy requirements. We provide a protection schemes for privacy-preserving speaker verification, where the system is able to perform verification without revealing personal information about the user. This thesis addresses privacy issues for speaker verification systems based on Gaussian Mixtures Models [Reynolds et al., 2000], i-vectors [Dehak et al., 2010] and x-vectors [Snyder

et al., 2017] as speaker representations.

Motivation 2: Although various privacy-preserving biometric systems have been proposed in the literature, most of the biometric protection schemes applied as countermeasures to improve the privacy and security issues degrade the biometric performance in terms of verification accuracy compared to the unprotected system. Therefore, there is a need for protection schemes that achieve privacy requirements while maintaining the biometric performance.

Objective 2: Providing biometric protection schemes that preserve the user's privacy while maintaining the biometric performance of the unprotected system in terms of verification accuracy.

Motivation 3: The evaluation of protection schemes proposed for different privacy-preserving biometric systems lacks a common and standardized assessment of privacy, security, and biometric performance. Most publications aren't taking into account all necessary aspects of a rigorous privacy and security evaluation.

Objective 3: Making a step forward towards the evaluation of privacy-preserving biometric systems on common and standardized assessments using public databases. The privacy will be evaluated according to the requirements described on the ISO/IEC 24745 for biometric information protection [ISO/IEC JTC1 SC27 Security Techniques, 2011]. Security will be analyzed according to the methodology proposed in [Rosenberger, 2018], where different attack scenarios are proposed to evaluate the robustness of biometric systems based on protection schemes. Related to objective 2, for a fair comparison, the biometric performance of the unprotected (baseline) and protected systems will be reported using common protocols and databases.

1.4 Outline of The Thesis

This dissertation is composed of seven chapters structured as follows:

– Chapter 1 introduces the privacy and security issues related to speaker verification systems and presents the motivation, objectives outline, and contributions of this thesis.

– Chapter 2 summarizes previous work related to the thesis topic.

– Chapter 3 presents the evaluation of the audio-visual recognition system developed during the H2020 European project SpeechXRays [4] against spoofing attacks based on a 3D talking head. We will demonstrate that the fusion of voice and face modalities can be a solution to improve the biometric performance but it is not sufficient to guarantee the security and privacy aspects of the user. This evaluation served as a motivation to develop privacy-preserving speaker verification systems.

– Chapter 4 describes the proposed biometric protection scheme for privacy-preserving speaker verification system based on Gaussian Mixture Models. The protection scheme includes two steps: first, the extraction of binary representation for the speaker derived from his/her GMM model and then the protection of the binary representation using a cancelable scheme named shuffling.

– Chapter 5 presents the proposed biometric protection scheme for privacy-preserving speaker verification system based on i-vectors. The protection scheme is also based on the binarization of the speaker's i-vector using a thresholding method and then its protection with the shuffling scheme.

– Chapter 6 presents the proposed biometric protection scheme for privacy-preserving speaker verification system based on x-vectors. For this scheme, we propose a novel binarization approach. This approach transforms the speaker's x-vector into a binary vector without a loss in biometric performance and makes it possible to control the dimension of the binary vector. Then, for the protection of the binary x-vector, we propose to combine the shuffling scheme with error correction code (ECC).

  Chapters 4,5 and 6 are structured as follows. We start by introducing the baseline (unprotected) speaker verification system. Then, we present the protection method proposed to develop the privacy-preserving speaker verification system. We describe the novel architecture and the steps required

[4] http://www.speechxrays.eu/

to ensure the user's privacy. The proposed system is then evaluated according to the requirements of biometric information protection [ISO/IEC JTC1 SC27 Security Techniques, 2011] in terms of biometric performance, revocability, irreversibility, and unlinkability. Also, the robustness against different attack scenarios is analyzed.

– Chapter 7 concludes ans summarizes the main results obtained in this thesis and outlining future work.

1.5 Research Contributions

The research contributions of this PhD thesis are the following:

1. Privacy-preserving speaker verification system based on Gaussian Mixtures models.

   MTIBAA, Aymen, PETROVSKA-DELACRÉTAZ, Dijana, et HAMIDA, Ahmed Ben. Cancelable speaker verification system based on binary Gaussian mixtures. In : 2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). IEEE, 2018. p. 1-6.

2. Privacy-preserving speaker verification system based on i-vectors.

   MTIBAA, Aymen, PETROVSKA-DELACRÉTAZ, Dijana, BOUDY, Jérôme, et al. Privacy-preserving speaker verification system based on binary I-vectors. IET Biometrics, 2021, vol. 10, no 3, p. 233-245.

3. Privacy-preserving speaker verification system based on x vectors. This contribution is described in chapter 6 but has not yet been published.

4. Participation in writing a survey about preserving privacy in speaker and speech characterisation.

   NAUTSCH, Andreas, JIMÉNEZ, Abelino, TREIBER, Amos, et al. Preserving privacy in speaker and speech characterisation. Computer Speech Language, 2019, vol. 58, p. 441-480.

5. Participation in the 2019 NIST Speaker Recognition Evaluation. The participation was described in chapter 9 of the IET book titled Voice Biometrics: Technology, trust, and security.

   HMANI, Mohamed Amine, MTIBAA, Aymen, PETROVSKA-DELACRTAZ, Dijana. Joining Forces of Voice and Facial Biometrics: a Case Study in the Scope of NIST SRE'19. In Voice Biometrics: Technology, trust and security (chapter 9). IET

6. This thesis is carried out in the context of two H2020 European projects, SpeechXRays [4] and EMPATHIC [5]. For the SpeechXRays project, I participated in the organisation of the SpeechXRays spoofing challenge [6]. Also, I contributed to the biometric evaluation of the audio-visual system that was developed and tested on three use-cases with 2000 users. This contribution is described in the following papers.

   MTIBAA, Aymen, HMANI, Mohamed Amine, PETROVSKA-DELACRÉTAZ, Dijana, et al. Methodologies of Audio-Visual Biometric Performance Evaluation for the H2020 SpeechXRays Project. In : 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). IEEE, 2020. p. 1-6.

   HMANI, Mohamed Amine, MTIBAA, Aymen, PETROVSKA-DELACRTAZ, Dijana, et al. Evaluation of the H2020 SpeechXRays project Cancelable Face System Under the Framework of ISO/IEC 24745: 2011. In : 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). IEEE, 2020. p. 1-6.

   SPANAKIS, Emmanouil G., PETROVSKA-DELACRÉTAZ, Dijana, BAUZOU, Claude, et al. Multi-Channel Biometrics for eHealth Combining Acoustic and Machine Vision Analysis of Speech, Lip Movement and Face: a Case Study. In : 2019 IEEE International Conference on Imaging Systems and Techniques (IST). IEEE, 2019. p. 1-6.

[5] http://www.empathic-project.eu/
[6] http://www.speechxrays.nipne.ro/spoofing-challenge

For the EMPATHIC project, I created the 3D-virtual coach designed to improve the independent Years of the Elderly. This work is described in the following paper.

TORRES, María Inés, OLASO, Javier Mikel, MONTENEGRO, César, et al. The empathic project: mid-term achievements. In : Proceedings of the 12th ACM International Conference on Pervasive Technologies Related to Assistive Environments. 2019. p. 629-638.

Chapter 2

Related Works

Contents

This chapter summarizes previous works related to this thesis. First, we present a brief review of existing speaker verification systems in Section 2.1. In Section 2.2, we present the vulnerabilities of speaker verification systems. Then, in Section 2.3, we review the state-of-the-art of existing works that address the privacy and security issues based on biometric information protection techniques. Finally, the summary and conclusion of this chapter are presented in Section 2.4.

2.1 Brief State of The Art of Speaker Verification Systems

Speaker verification (SV) is the process of accepting or rejecting the claimed identity of a speaker, based on his/her voice characteristics (features) extracted from recorded samples. The features could be extracted using mel-frequency cepstral coefficients (MFCC), filter bank (FBANK), linear predictive cepstral coefficients (LPCC), perceptual linear prediction (PLP) [Lawson et al., 2011], or directly from the raw waveforms using neural network [Palaz et al., 2015], [Jung et al., 2018]. SV can operate on two scenarios: text-dependent and text-independent. The text-dependent scenario requires that the probe spoken text be the same as the enrollment. In contrast to that, in the text-independent scenario, no prior constraints are considered for the spoken phrases by the speaker during the verification.

One of the first successful approaches for speaker verification is the Gaussian Mixture Model (GMM) [Reynolds et al., 2000]. In this approach, the features are modeled using a GMM by adapting the enrollment voice samples to a universal background model (UBM) that represents the distribution of the acoustic features of a large population of speakers. In verification, the likelihood ratio of the probe features given the enrollment GMM and the UBM is computed to take the verification decision.

Dehak *et al.* [Dehak et al., 2010] introduced SV based i-vector, where features are represented by a low-dimensional fixed-length vector. From a sequence of feature vectors, e.g. MFCC, sufficient statistics are collected and represented by Baum-Welch statistics obtained with respect to a UBM. Then, these statistics

are converted into a low-dimensional representation know as i-vector. For verification, the similarity between i-vectors is measured by simple cosine similarity or using a more elaborate Bayesian model such as Probabilistic Linear Discriminant Analysis (PLDA) [Kenny, 2010], [Garcia-Romero and Espy-Wilson, 2011].

Lei *et al.* [Lei et al., 2014] proposed a framework in which the sufficient statistics for the i-vector are driven by a deep neural network (DNN) trained for automatic speech recognition (ASR). The DNN is used to enhance phonetic modeling in the i-vector: either posteriors from the DNN replace those from a Gaussian mixture model (GMM) [Kenny et al., 2014], or bottleneck features are extracted from the DNN and concatenated with acoustic features [McLaren et al., 2015].

Recent approaches proposed to replace the GMM-UBM and i-vector models by speaker representation based only on deep networks. Variani *et al.* [Variani et al., 2014] propose a DNN based speaker verification for text-dependent task. The DNN was first trained to classify the speakers at the frame-level on the sentence "Ok Google". Then, it is used to extract a novel representation of features from the last hidden layer. The average of these novel speaker features is taken as the speaker representation known as d-vector. Based on this approach, in [Heigold et al., 2016] an end-to-end speaker verification system was presented. This system maps the enrollment and probe utterances directly to a single score for verification.

Snyder *et al.* [Snyder et al., 2016] propose an end-to-end text-independent speaker verification system. The system is based on DNN trained to discriminate between same speaker and different speaker. It takes as input a variable-length utterance and maps it to a speaker embedding by aggregating the frame-level representations using a pooling layer. In [Snyder et al., 2017], instead of training the system to separate same-speaker and different speaker pairs, the DNN learns to classify a set of training speakers using categorical cross-entropy loss. The DNN consists of layers that operate on speech frames, a statistics pooling layer that aggregates over the frame-level representations, additional layers that operate at the segment level, and finally a softmax output layer. Speaker representation known as x-vector embedding is then extracted from any layer after the statistics pooling layer. The performance of this system was improved in [Snyder et al., 2018] by augmenting the x-vector training data with additive and convolutional noise. Reported results show that x-vector embeddings outperformed i-vectors in

terms of biometric performance.

For this thesis, we address speaker verification systems based on GMM, i-vector, and x-vector. More description of these approaches will be provided in Chapter 4, 5 and 6 respectively.

2.2 Vulnerabilities of Speaker Verification Systems

In recent years, speaker verification systems have shown an improvement in both accuracy performance and their practical use. Public acceptance, availability, and the low prices of microphones have promoted the integration of such technologies into our daily lives. However, this evolution has led to critical issues related to the security and the privacy of the user.

The process of speaker verification system requires that the access system stores the speakers' representations and has access to the recordings or features derived from the speakers' voices during the test. This process poses threats to privacy and security. Moreover, there are various vulnerabilities through which a non-target user can attack the biometric systems as shown in Figure 2.1. Regarding speaker verification systems, the most vulnerable points in such systems are at levels 1 and 2 correspondings to presentation attacks at sensor level and at acquisition level before the signal processing. The voice samples of a given user can be easily collected through face-to-face recording, telephone conversation, or compromised databases and then used to spoof the system or to extract personal information. Also, with advanced technologies in speech synthesis or voice conversion, we can generate the voice of the target user and manipulate it to gain unauthorized access. Various methods are proposed in the literature for the voice impersonation attacks [Sahidullah et al., 2019] which are classified as human-based voice impersonation, replay-based attacks, speech synthesis, and voice conversion attacks.

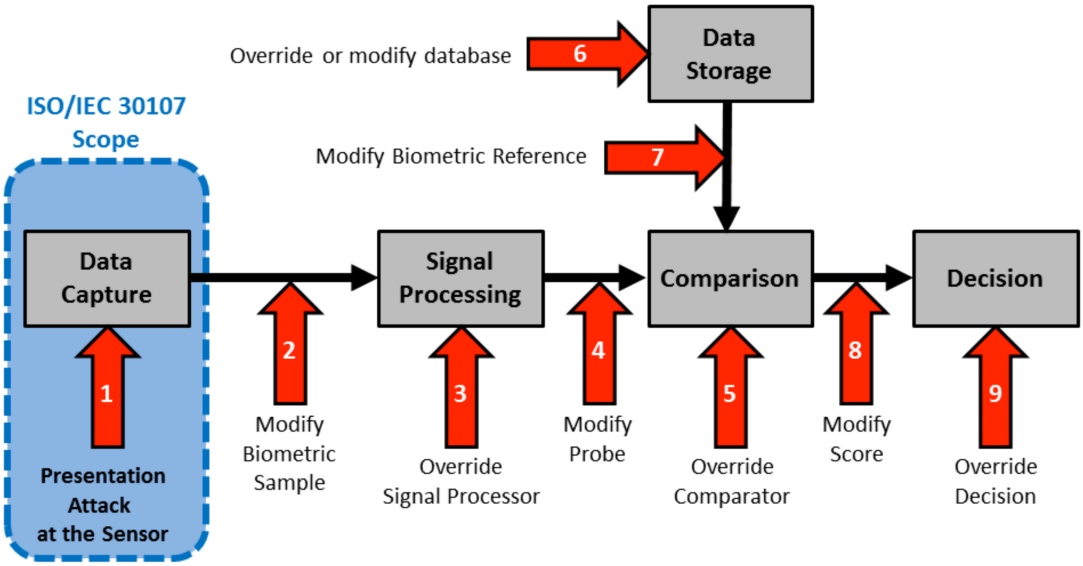


Figure 2.1: Presentation attacks on biometric systems [ISO/IEC 30107, 2017].

2.2.1 Impersonation attack

In this attack, the non-target user modifies his/her voice to imitate the target user's voice and tries to spoof the system.

In [Lau et al., 2005], the authors studied the voice mimicry with the GMM-based speaker verification system. Professional and non-professional imitators were asked to imitate a selected speaker from the YOHO database. Experiments on 138 speakers in the YOHO database and six participant who played a role as imitators showed a fact that professional imitators could successfully attack the system and that non-professional users could have a good chance if they know their closest speaker in the database. At the threshold correspond to FAR=0 for the baseline system, a FAR=60% was reported with the professional imitators and FAR=20% with the non-professional imitators.

Farrus *et al.* [Farrús Cabeceran et al., 2010] conducted experiments with professional imitators to perform impersonation attacks on speaker identification systems based on prosodic features. Two male professional imitators attempt to mimic the voice characteristics of five well-known male politicians. Experiments show that an increase from an identification error rate of 5% for target speakers against the impersonator's natural voice to an identification error rate of 22% for target speakers

against the impersonator's modified voice.

Panjwani *et al.* [Panjwani and Prakash, 2014] involve crowdsourcing method to find impersonators that are used to perform impersonation attacks on text-independent GMM-based speaker verification. The experiments are conducted using a database collected with 53 male Indian voices. Results showed that crowdsourcing method can identify non-professional impersonators with a high acceptance rate. From a pool of 176 candidates, they identified six impersonators with an overall false acceptance rate of 44% compared to 2.31% for the baseline system. This demonstrates that naive, untrained users have the potential to carry out impersonation attacks against voice-based systems.

The GMM-UBM, i-vector with cosine scoring, and i-vector with PLDA scoring based speaker verification systems were evaluated against impersonation attacks in [Hautamäki et al., 2015]. A speech database containing the voice of eight well-known Finnish public figures was used for this evaluation. Results show that the impersonation attack decreased the EER for GMM-UBM from 10.83% to 10.31%, while for i-vector systems the EER increased from 6.80% to 13.76% and from 4.36% to 7.38%.

Mandalapu *et al.* [Mandalapu et al., 2021] analyzed the vulnerabilities of i-vector and x-vector speaker verification systems using a database collected for voice impersonation attack. The speakers in the database include politicians and actors. The bona fide speeches are taken from the interview videos of the target speakers. The impersonation speeches are collected from YouTube videos of television shows and performances by mimicry artists ranging from amateurs to professionals. The evaluations show that for the speaker verification system based on i-vector, the EER increased from 5.3% (baseline) to 12.9% (impersonation attack) and for the x-vector system the EER increased from 3.8% (baseline) to 11.10% (impersonation attack).

2.2.2 Replay attack

This attack consists of a non-target user trying to use a pre-recorded voice of a target user to spoof the system. This attack presents one of the main weaknesses of the SV systems, especially in the text-independent scenario. The voice samples

of a given user can be easily collected through face-to-face recording, telephone conversation, or public video which poses privacy and security threats.

Villalba *et al.* [Villalba and Lleida, 2010], [Villalba and Lleida, 2011] studied the vulnerabilities of speaker verification system-based joint factor analysis in the case of the text-independent scenario. The system was evaluated using voice samples recorded through a far-field microphone and then replayed using a mobile phone (the studies involved five speakers). Experimental results show that EER of 0.71% is obtained using the non spoofing trials and if the EER operating point is taken as the decision threshold, the system accepts 68% of the spoofing trials.

Ergunay *et al.* [Ergünay et al., 2015] present an audio-visual spoofing database for replay attacks collected using a low and high-quality microphone from phones and laptops. Using this database, the vulnerability of speaker verification-based i-vector was evaluated against replay attacks. Biometric performance of the baseline i-vector was reported with an EER equal to 6.9% for males and 17.5% for females. When applied the replayed attack, at the EER operating point, the FAR increased to 77.4% and 69.4% for males and females, respectively.

2.2.3 Speech synthesis attack

Speech synthesis, also known as text-to-speech (TTS), is a method for producing a speech signal from a given text. Due to different methods as unit selection [Hunt and Black, 1996], [Senior and Fructuoso, 2016], statistical parametric [Zen et al., 2009], and Deep neural speech generation [Kaneko et al., 2017], [Mehri et al., 2016], [Shen et al., 2018], it was possible that a non-target user synthesizes a natural voice similar to a target user and gains unauthorized access to the system.

In [Ergünay et al., 2015] the vulnerability of speaker verification-based i-vector was evaluated against speech synthesis attacks using the AVspoof database[1]. The speech synthesis attacks were based on statistical parametric speech synthesis (SPSS) [Zen et al., 2009]. Hidden Markov model-based speech synthesis technique [Yoshimura et al., 1999] was used to produce high-quality synthetic speech. Using the synthetic speech, at the EER operating point, a FAR equal to 94.1% was reported compared to 4.9% for the baseline system.

[1]https://www.idiap.ch/en/dataset/avspoof

In [De Leon et al., 2012], the vulnerability of speaker verification systems based on GMM-UBM and SVM using GMM supervectors against synthetic speech was evaluated. HMM-based text to speech synthesizer was used to generate synthesize voice of target users. Using the Wall Street Journal (WSJ) corpus, they have shown that over 81% of synthetic speech signals compared to a target user are accepted which poses a potential security issue.

Cai *et al.* [Cai et al., 2018] investigate the ability of generative adversarial network (GAN) to synthesize spoofing attacks on speaker identification systems based on Mel-Spectrogram and convolution neural networks. They show that adversarial samples generated with GAN networks are successful in performing targeted and untargeted adversarial attacks.

2.2.4 Voice conversation attack

Voice conversion (VC) has become one of the most easily accessible techniques with the available applications to carry out spoofing attacks. It aims to modify the attacker's voice to sound like it was pronounced by the target speaker [Wu and Li, 2013]. It presents a threat to both text-dependent and text-independent speaker verification systems.

In [Matrouf et al., 2006], the vulnerability to voice conversion attack was evaluated for speaker verification system based on GMM-UBM. The voice conversion was performed by mapping the attacker's vocal tract information towards that of the target user using the frequency warping technique. Experimental results reported on NIST SRE 2005 database show that the EER degrades from 10% to 60% when attacker voice samples are compared to the target users.

In [Alegre et al., 2012], the vulnerability of text-independent speaker verification systems based on GMM-UBM and JFA were evaluated against voice conversion attacks. Experimental results on the male test set of NIST SRE 2005 show that the EER is increased from 8.5% and 4.8% to 32.6% and 24.8% for GMM-UBM and JFA systems, respectively.

Kinnunen *et al.* [Kinnunen et al., 2012] studied the vulnerability of text-independent speaker verification systems (GMM, JFA) against voice conversion attacks using telephone speech. A voice conversion system was implemented with

two types of features and nonparallel frame alignment methods. Experiment results on a subset of NIST SRE 2006 corpus indicate that the FAR of the most robust JFA system increased from 3% to over 17%.

In order to avoid the vulnerabilities related to spoofing attacks, four editions of automatic speaker verification and spoofing countermeasures challenges ASVspoof [2] have been organized. While the first edition in 2013 [Evans et al., 2013] was targeted mainly at increasing awareness of the spoofing problem, the 2015 edition [Wu et al., 2015] included the first challenge on the topic, accompanied by commonly defined evaluation data, metrics, and protocols. The task in ASVspoof 2015 was to design countermeasure solutions capable of discriminating between genuine speech and spoofed speech produced using either text-to-speech or voice conversion systems. The ASVspoof 2017 challenge [Kinnunen et al., 2017] focused on the design of countermeasures aimed at detecting replay spoofing attacks and the 2019 edition [Todisco et al., 2019] focused on countermeasures for three attack types, namely those stemming from TTS, VC, and replay spoofing attacks. The ASVspoof 2021 [Yamagishi et al., 2021] was the 4th edition where the goal was to develop countermeasures capable of discriminating between bona fide and spoofed or deepfake speech.

In addition, biometric information protection schemes were also proposed as a countermeasure to prevent the success of the spoofing attacks and thereby enhance the privacy provided by SV biometric systems. In the following section, we present the existing works related to the development of biometric information protection schemes for speaker verification systems

2.3 Biometric Information Protection Schemes for Speaker Verification Systems

Various researches have contributed to the development of privacy-preserving biometric systems [Kumar et al., 2020], [Rathgeb and Uhl, 2011] and proposed bio-

[2] https://www.asvspoof.org/

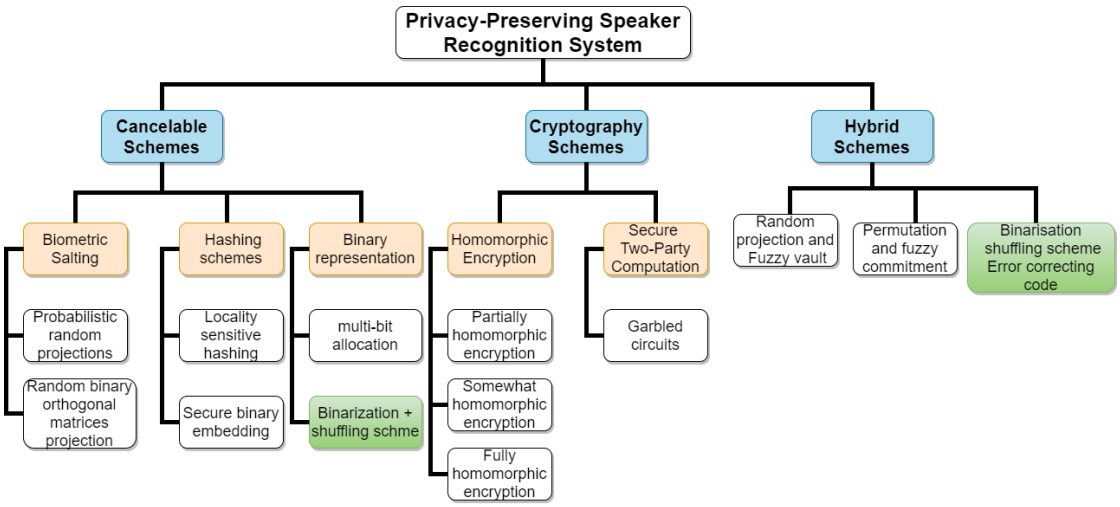


Figure 2.2: Classification of privacy-preserving schemes for speaker recognition systems. The categories to which the schemes proposed in the thesis belong are highlighted in green.

metric protection schemes devoted to preserving the privacy of biometric systems based on face, iris, and fingerprint modalities where biometric references are represented with templates. However, such schemes can not be applied for some speaker verification systems where the user is represented with models rather than templates. As example, for speaker verification system based on Gaussian mixture models, we need to develop a protection schemes to protect the GMM models.

A recent survey of existing biometric information protection schemes that address the privacy-preserving in the context of speaker recognition systems was presented in [Nautsch et al., 2019]. In Figure 2.2, we present a classification of these systems based on the schemes used to achieve the privacy requirements. We classified the schemes into three categories, (i) cryptography-based schemes; (ii) cancelable based schemes; and (iii) hybrid based schemes. In the following subsections, summaries of related works related to the three categories are presented.

2.3.1 Cryptography based schemes

For cryptography based schemes, techniques such as Homomorphic Encryption (HE) and Secure Two-Party Computation (STPC) are used to protect the biometric data by encrypting the biometric reference and the biometric comparison

is carried out in the encrypted domain.

2.3.1.1 Privacy-preserving based on homomorphic encryption

Homomorphic Encryption schemes make it possible to perform computations on ciphertexts and generate encrypted results without requiring any decryption. The decryption of these results to plaintext [3] corresponds to the results of the operations carried out on the original plaintext. The structure of the plaintext space is preserved in the ciphertext space for additions and/or multiplications of plaintext data under encryption [Acar et al., 2018]. Therefore, combining such encryption techniques with biometric verification systems allows achieving privacy requirements while maintaining the biometric performance. Figure 2.3 presents a general pipeline of privacy-preserving biometric systems based on HE.

HE is divided into three types, Fully Homomorphic Encryption (FHE) that allows unlimited additions and multiplications at the cost of an increased computational load [Gentry, 2009], Somewhat Homomorphic Encryption (SHE) that has a fixed limit of multiplications to speed up their execution, and Partially Ho-

[3]This usually refers to data that is transmitted or stored unencrypted ("in clear").

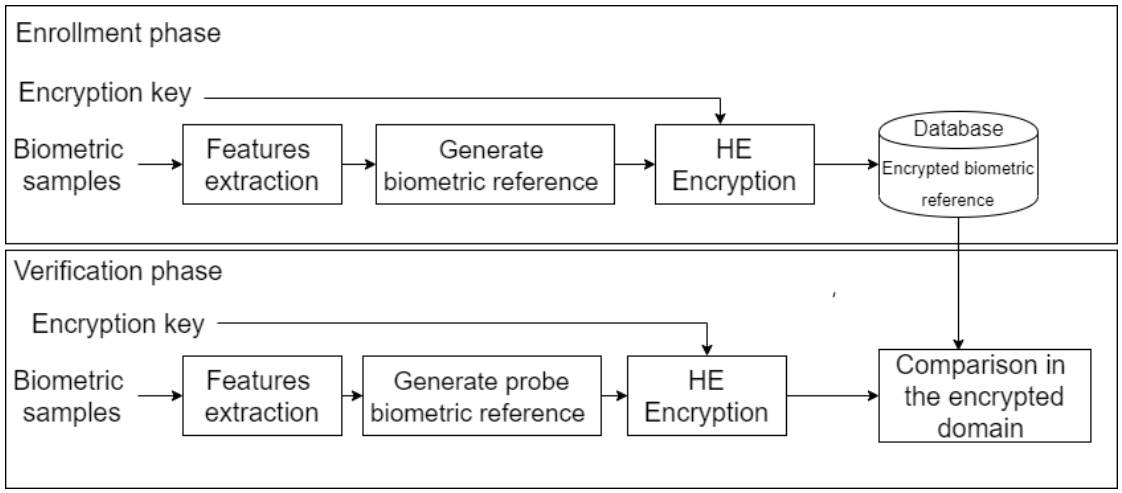


Figure 2.3: Privacy-preserving based on Homomorphic encrytion. The user's biometric reference is encrypted during the enrollment. During the verification, the probe features or biometric references are encrypted and the comparison is performed in the encrypted domain using Homomorphic operations.

momorphic Encryption (PHE) that supports either additions or multiplications, hence, they are only partially homomorphic.

Partially homomorphic schemes such as Pallier encryption [Paillier, 1999] was adapted to preserve privacy for speaker verification system based on GMM-UBM and i-vector. Pathak *et al.* [Pathak and Raj, 2011], [Pathak and Raj, 2012a] developed a privacy-preserving protocol for speaker verification systems based on GMM using Pallier encryption and STPC protocols [Yao, 1982]. During the enrollment phase, the user has the enrollment samples and both encryption and decryption keys. On the other side, the system has the UBM and the encryption key. To start, the system sends the UBM to the user in plaintext to derive his/her GMM model. Then, the user encrypts the GMM model with his/her key and sends it to the system. In the end, the system has only the encrypted GMM models. During the verification phase, the user encrypts the features of the probe voice sample and sends it to the system that computes the log-likelihoods for encrypted frames and encrypted mixture components of the GMM model in the encrypted domain using homomorphic operations. The comparison score is obtained from the encrypted log-likelihoods score using logsum protocol, which requires additional communication between the system and the user. Based on this protocol, the verification is performed without that the system observes the features provided by the user and the user does not observe the models stored in the system. This approach achieves privacy requirements while maintaining the biometric performance of the baseline system. However, the limitation of this approach is the huge computational overhead compared to the baseline speaker verification system based on GMM-UBM due to the large amount of time required to perform operations in the encrypted domain.

In [Nautsch et al., 2018], homomorphic encryption based on Pallier cryptosystem was also used as a privacy-preserving solution for speaker verification based on i-vector using cosine or PLDA as back-end scoring. The solution is based on two-colluding servers named $DB_{controller}$ and $AS_{operator}$. During the enrollment phase, the user's i-vector is encrypted using the public key of the authentication server $AS_{operator}$ and stored in the database server $DB_{controller}$. During the verification phase, the user extracts probe i-vector, and the enrollment encrypted i-vector is sent to the user device for the process of verification. The comparison

score is computed in the encrypted domain using homomorphic operations. Then, the encrypted score is sent to $AS_{operator}$, which decrypts the score and takes the verification decision. Experimental results show that this solution preserves the verification performance and achieves privacy requirements. However, the use of HE results in a high communication and computation overhead that makes this solution impractical when considering computationally limited devices. For i-vectors of dimension 600, the computations require 203 milliseconds per comparison when only subject data is encrypted using cosine as a back-end scoring, 423 ms using unprotected PLDA model, and 2171 seconds per comparison when both subject data and PLDA model parameters are encrypted. Moreover, this solution is vulnerable in terms of security to authentication by a malicious user that can compromise the system by just sending the encryption of an accepting score to the $AS_{operator}$.

2.3.1.2 Privacy-preserving based on secure two-party computation

Secure two-party computation allows two parties to interactively compute any function in a secure manner without revealing the plaintext. Therefore, it was exploited as a solution to develop privacy-preserving speaker verification systems.

Yao's Garbled Circuit (GC) [Yao, 1982] has been employed to preserve the privacy of GMM-UBM speaker verification system. Portélo *et al.* [Portêlo et al., 2014a] reformulated the GMM-based speaker verification by performing the required operations like the scalar product and logsum operations using the Garbled Circuit. The proposed protocol assumes that the user is responsible for generating the GCs and the system is responsible of evaluating them and deciding on whether or not to authenticate the user. Experimental results show that the proposed system achieves a biometric performance close to the unprotected system and guarantees that each of the participants in the protocol does not reveal his/her private information to others. Also, in terms of execution time, this solution is faster than HE-based schemes, but it scales linearly with the number of GMM components. A drawback of this scheme is the fact that the verification system has the user-specific GMM model in plaintext which represents a privacy leak because the system is in possession of a characterization of the user's voice given by

the parameters of the GMM.

Privacy-preserving of Hidden Markov model (HMM) was also treated in [Aliasgari et al., 2017] by storing secret shares among multiple servers using a technique known as outsourced secure multi-party computation (SMPC). The solution uses floating-point arithmetic, which allows to achieve privacy and security guarantees while maintaining reasonable performance. Also, SMPC significantly decreases workload compared to HE.

Similar to the previous scheme, Treiber *et al.* [Treiber et al., 2019] proposed a privacy-preserving i-vector speaker verification system based on a mix of different STPC protocols. This solution achieves biometric information protection requirements while maintaining biometric performance. Also, in contrast to the solution based on HE in [Nautsch et al., 2018], the verification using PLDA as back-end scoring is computed in about half a second, and a few milliseconds using cosine as back-end. However, it involves multiple rounds of interaction and communications between parties involved in the secure computation.

Recently, Nautsch *et al.* [Nautsch et al., 2019] addressed the issue of computational overhead. They proposed a solution that enables privacy-preserving of i-vector speaker verification system with cohort score normalisation using probabilistic linear discriminant analysis comparisons. The solution proposes a cohort pruning scheme based on secure multi-party computation that operates with binary voice representations to reduce the computation time for biometric comparisons in the encrypted domain.

2.3.1.3 Summary of cryptography based schemes

The use of HE encryption allows to preserve privacy while maintaining the biometric performance obtained with the unprotected system. However, the size of the encrypted data and the huge number of operations required in the encrypted domain, result in overheads of computation and communication, which slows down the verification process. HE-based solutions rely on noise to hide the plaintext. This noise grows during processing speech data in the encrypted domain due to the homomorphic operations (addition, multiplication) required. As a result, the calculation will be performed with larger data than the actual plaintext and the

noise will eventually overflow. Hence, an expensive operation named bootstrapping [Gentry, 2009] is introduced to reduce it, making the computational overhead too heavy. Thus, the integration of these schemes while keeping verification time low enough for real-time applications is very challenging, especially when considering computationally limited devices such as mobile phones.

STPC and SMPC protocols were also applied to the privacy preservation of speaker verification systems. These protocols show an improvement in the verification execution time compared to HE-based solution. However, it involves multiple rounds of interaction and communications between parties involved in the secure computation, and privacy is achieved when assuming that the different parties do not collude.

2.3.2 Cancelable biometrics based schemes

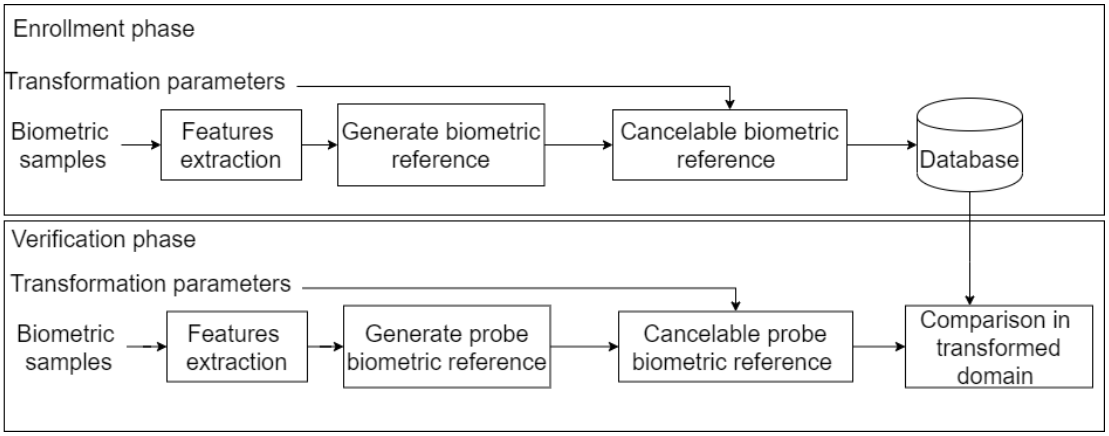


Figure 2.4: Privacy-preserving based on cancelable biometrics. The user's model or template is distorted with transformation parameters during the enrollment and stored in the database. During the verification, the probe biometric reference is transformed with the same transformation parameters used during the enrollment, and the comparison is performed in a transformed domain. For cancelable biometric, the transformation could be applied either to biometric references (template /model) or to the features.

As an alternative to the aforementioned biometric information protection schemes based on homomorphic encryption and secure two-party computation, cancelable

biometric was proposed as a solution to develop privacy-preserving speaker verification systems. As shown in Figure 2.4, for cancelable biometrics, the user transforms the biometric reference before sending it to the access server. The server receives only a protected format of the original biometric reference and during the verification, the biometric comparison is performed in the transformed domain using the protected biometric references. Regarding the existing works on cancelable speaker verification systems, we classified the cancelable schemes into three categories:

(i) Based on biometric salting schemes: where privacy-preserving is achieved by combining an Auxiliary Data (AD) with biometric features or reference to derive a protected version of the biometric reference.

(ii) Based hashing schemes: where privacy-preserving is achieved using hashing techniques.

(iii) Based on binary speaker representation: where privacy-preserving is first achieved by transforming the speaker model into a binary representation, then cancelable schemes (i) or (ii) are applied to transform the binary representation.

In the following paragraphs, we present a description of the existing cancelable schemes for each category.

2.3.2.1 Privacy-preserving based on biometric salting schemes

Chee *et al.* [Chee et al., 2018] proposed a cancelable scheme, named Random Binary Orthogonal Matrices Projection (RBOMP), to protect speaker verification system based on i-vector. The RBOMP scheme projects the i-vector using random binary orthogonal matrices from linear space to ordinal space and records the discrete values. In order to achieve the irreversibility requirement, a non-invertible function (prime factorization) is used to protect the projected i-vector with the help of a user-specific token. The protected system was evaluated using NIST SRE 2010 extended condition 5 (tel-to-tel) female part. The cancelable scheme shows good resistance against irreversibility and attack-via-record multiplicity. However, the biometric performance of the cancelable system (EER=3.43%) degrades compared to the baseline i-vector system (EER=1.67%).

Teoh and Chong [Teoh and Chong, 2010] provided a cancelable GMM speaker verification system based on probabilistic random projections [Teoh and Yuang, 2007]. This scheme protects the speaker's model by hiding the features through a random subspace projection process and its parameters are stored in a subject-specific key. This method achieves the revocability and unlinkability requirements and it is shown that the cancelable system maintains the biometric performance.

2.3.2.2 Privacy-preserving based hashing schemes

Hashing techniques were also used to protect speaker verification systems in [Pathak and Raj, 2012b]. The idea was to transform the speaker verification task into string comparison. For this, the voice samples provided by the user are represented using supervectors features [Campbell et al., 2006] and passed through a Locality Sensitive Hashing (LSH) to transform them into bit strings. Then, the bit strings are converted into obfuscated strings by applying a cryptographic hash function. The biometric comparisons are then performed by matching the hashed bit strings derived from the enrollment and the probe samples. This approach performs speaker verification without revealing the voice samples provided by the user to the system. Moreover, cryptographic hash functions are faster to compute adding a small overhead compared to the overhead of the secure multiparty computation approaches using homomorphic encryption proposed in [Pathak and Raj, 2012a]. However, while HE preserves the biomectric performance of the unprotected system, the LSH transformation shows a degradation in biometric performance with an EER=11.86% on the YOHO database.

Jimenez and Raj [Jiménez and Raj, 2017] proposed a two-factors transformation to perform speaker verification based on GMM supervectors without revealing user's biometric information to the system. This transformation is based on distance-preserving hashing. By combining a user-specific key with the voice features, the transformation allows to detect if the distance between the transformed features is smaller than a verification threshold without revealing the original features. Experimental results show that the proposed transformation improves the biometric performance. However, in case the user specific-key is compromised a degradation in performance was reported.

Portélo *et al.* [Portêlo et al., 2014b] proposed a cancelable i-vector system that performs speaker verification without exposing voice samples or models to the system. The cancelable scheme is based on the transformation of the speaker's i-vector (float) to bit sequences using the Secure Binary Embeddings (SBE) [Boufounos and Rane, 2011]. Then a support vector machine (SVM) classifier is modified to work with the Hamming distance between the SBE hashes of i-vectors. During the enrollment, the user computes the SBE hashes from his/her enrollment i-vector and transmits it to the system. The parameters used by the SBE are considered as the user's private keys. Also, the system trains an SVM with the obtained SBE hashes. During the verification phase, the user computes the SBE hash for the probe i-vector and transmits it to the system, which classifies it using the trained SVM. Based on this protocol, the system does not observe the user's i-vector in plain-text. Results reported with hashed i-vectors show that the speaker verification performance depends on parameters fixed for the SBE, including the number of bits in the hashed i-vector and the amount of data leakage from the speaker representation. With the best configuration of these parameters to achieve high privacy, the proposed cancelable system does not maintain biometric performance compared with the unprotected i-vector system. Besides, this scheme was not evaluated according to the biometric information protection requirements such as irreversibility and unlinkability. There is no guarantee that a non-target user is not able to infer information about the non-protected i-vectors when he succeeds to obtain the secrets parameters of the SBE.

2.3.2.3 Privacy-preserving based on binary speaker representation

Regarding the biometric information protection schemes used for the protection of facial, iris and fingerprint biometric recognition systems, most of these schemes require a binary representation of features or templates. In order to exploit these protection schemes to protect speaker verification systems, binary representations developed originally for biometric speaker verification or diarization were used for privacy-preserving.

Paulini *et al.* [Paulini et al., 2016] proposed a binarization method for voice biometric features known as multi-bit allocation. It is designed to extract discrim-

inative compact binary feature vectors to be applied in a voice biometric template protection scheme. The binarization acts over GMM super-vectors estimated over MFCC features. The feature space is divided into intervals, which are encoded with multiple bits using a Gray code. Experimental results show that the binary representation of voice features causes a negligible decrease in biometric performance compared to the baseline system.

Billeb *et al.* [Billeb et al., 2015] proposed a binarization method based on GMM-UBM, that is used to extract high-entropy binary voice template from speaker models. Speaker binary templates are then protected with a fuzzy commitment scheme [Juels and Wattenberg, 1999], which combines techniques from the area of error correcting codes and cryptography. Experimental evaluation has shown that the system achieves privacy requirements. However, the biometric performance degrades due to the binarization process.

Another binarization technique refereed as binary key speaker modeling was also proposeịn [Anguera and Bonastre, 2010]. This technique is designed to represent a sequence of acoustic features (MFCC) by a novel vector composed of binary values. The binarization process is based on three main blocks. The first block corresponds to the training of a generator model of N Gaussian components that are optimized to highlight speaker discriminant aspects. N represents the dimension of the binary vector. The second block corresponds to the extraction of the binary representation given a voice utterance as input. It is generally done in two steps. First, an accumulative vector V_c with dimension N is initialized to 0, and the likelihoods for each acoustic frame in the utterance are computed given each of the generator model Gaussian components. Then, the top Gaussians with the highest likelihood values are selected. An initial feature-level binarization is obtained by setting to 1 the bits in V_c corresponding to the positions of the top-scoring Gaussians. This process projects acoustic frame from the feature space into the space of the generator model Gaussians and keeps components with the highest impact. When all frames have been processed, each bit in V_c contains the relative importance of each Gaussian component in modeling the voice utterance given as input. In the second step, the final binary representation is obtained by setting to 1 in the binary vector, the bits corresponding to the top positions in the accumulative vector. Finally, a third block defines the distance between two

Table 2.1: Summary of cancelable biometric schemes for speaker verification systems.

Cancelable schemes	**Database**	**Speaker model**	**Baseline EER% before protection**	**Best EER% after protection**
Probabilistic Random Projection [Teoh and Chong, 2010]	Text-independent YOHO	GMM-UBM	5.37	0.27
Random binary Orthogonal Projection [Chee et al., 2018]	Chinese Mandarin digit corpus	i-vector	3.81	7.01
	Text-independent NIST SRE-2010		1.67	3.43
Secure Binary Embedding [Portêlo et al., 2014b]	Text-independent YOHO	GMM supervector	0.25	1.32
		i-vector	0.11	5.55
Locality Sensitive Hashing [Pathak and Raj, 2012b]	Text-independent YOHO	GMM supervector	-	11.8
Mullti-bit allocation [Paulini et al., 2016]	Text-independent digit corpus	GMM-UBM	3.4	3.56
Binarization + Fuzzy Commitment [Billeb et al., 2015]	Text-independent digit corpus	GMM-UBM	3.4	5.42

utterances by comparing the binary representations using similarity scores. This binarization technique was used for speaker recognition task in [Bonastre et al., 2011] and for speaker diarization in [Anguera and Bonastre, 2011] and [Delgado et al., 2015]. This technique will be also used in this thesis, to develop a privacy-preserving speaker verification system based on GMM.

Li *et al.* [Li et al., 2016] investigated the use of binary embeddings for speaker recognition. They studied two binarization approaches, one is based on LSH and the other is based on Hamming distance learning to transform i-vectors to binary vectors. Evaluations show that binary speaker embeddings deliver competitive

results on speaker recognition and reduce the computation cost.

2.3.2.4 Summary of cancelable biometrics

From the above-cited research, it is difficult to establish a fair comparison between the described schemes. Schemes were evaluated using different databases under different scenarios since a common and standardized evaluation of cancelable biometric is missed.

Regarding privacy requirements, revocability is preserved by combining the biometric information with a user-specific key, and different cancelable biometric references can be generated from the same biometric sample using different keys. As shown in Table 2.1, the limitation of some cancelable schemes is the degradation of biometric performance due to the modification or the loss of biometric information caused by the transformation schemes.

Cancelable biometric is based on two-factors, the biometric data, and the user-specific key. Therefore, the biometric performance of the cancelable system in case these factors are compromised should be reported. However, the majority of the described cancelable schemes do not evaluate the system in such scenarios. Moreover, unlinkability and irreversibility were not evaluated for most cancelable schemes. Compared to cryptography-based schemes, cancelable biometric preserves the execution time close to the unprotected system which makes it practical for real-time applications.

2.3.3 Hybrid schemes

Hybrid schemes consist of combining two or more schemes to generate protected speaker verification system as the combination of cryptography and cancelable schemes. Zhu *et al.* [Zhu et al., 2012] proposed a hybrid scheme based on the combination of random projection transformation with fuzzy Vault [Juels and Sudan, 2006] to generate protected speaker model. Experimental results show that this scheme preserves the biometric performance. However, privacy requirements such as unlinkability and irreversibility are not analyzed.

Inthavisas and Lopresti [Inthavisas and Lopresti, 2012] proposed a secure authentication system based on the combination of protected voice biometric tem-

plate and password using fuzzy commitment scheme [Juels and Wattenberg, 1999]. The system consists of three steps. In the first step, dynamic time warping template extracted from voice features is transformed using a password. Then the transformed template is mapped into a binary string. In the second step, the binary string is permuted using a password to avoid that an attacker predicts the correct password if the biometrics data are compromised. In the third step, the protected binary string and a cryptographic key are hidden using a fuzzy commitment framework. Experimental results show that the proposed system maintains the verification performance if the biometrics and passwords are not compromised simultaneously.

2.4 Chapter Summary and Conclusions

In this chapter, we have summarised the main works related to this thesis. First, we have started with a brief description of state-of-the-art speaker verification systems. Then, we have presented the main vulnerabilities of SV systems in terms of privacy and security. Next, we have summarized the existing works on the development of privacy-preserving speaker verification systems.

We have observed that systems based on cryptography schemes achieve privacy requirements while maintaining biometric performance. However, these schemes result in huge computational overhead. For cancelable biometrics, we have observed that most of the existing schemes degrade the biometric performance compared to the unprotected system. Besides, the complete set of privacy requirements to be validated according to the ISO/IEC 24745 are not evaluated for most of the proposed schemes.

For this thesis, we are focusing on cancelable biometrics. The objective is to propose privacy-preserving speaker verification systems based on cancelable schemes that achieve the privacy requirements of ISO/IEC 24745 while maintaining the biometric performance of the unprotected system.

Chapter 3

3D Talking Head Generation for Spoofing an Audio-Visual Biometric Recognition System

Contents

Multi-modal biometric as audio-visual biometric systems are used as a solution to ensure secure authentication. However, such systems threaten users' privacy, who are asked to provide an increased amount of sensible information. In this chapter, we show that the fusion of voice and face are not sufficient to guarantee security and privacy. We present the evaluation of the audio-visual recognition system developed during the H2020 European project SpeechXRays [1] against spoofing attacks based on a 3D talking head. This system proposes the fusion of voice and face as a solution to improve the robustness of the system. However, we will show that an adversary could use a voice recording and a 2D image of a target user to create an animated 3D facial model able to spoof the recognition system. Therefore, privacy-preserving schemes could be a solution to improve the privacy and the security.

3.1 Generation of 3D Talking Head

In this section, by exploiting a 2D image of the target user, we present how to create a 3D talking head. For that, we use facial animation tools that give the possibility of producing a 3D facial model that could be animated with voice recordings.

3.1.1 Creation of the 3D facial model from 2D image

CrazyTalk[2] facial animation software was used to generate a 3D head from the 2D facial image. The creation of the 3D face model is based on the adaptation of a generic 3D head mesh with the target 2D face image. As shown in Figure 3.1, we started by loading a frontal face image of the target user in CrazyTalk. Then, we manually identified 13 facial fitting points of the user's face. These points capture the pose, shape, and expressions of the user. Setting the fitting points to the correct positions on the face is required in order to have a better resemblance of the final 3D head to the original 2D image. Once the face landmarks points are fixed, we choose a head shape and generate the 3D facial model. Finally, some manual adjustments are required to have the correct appearance. These adjust-

[1] http://www.speechxrays.eu/
[2] https://www.reallusion.com/crazytalk/

ments include specifications such as cleaning the eyes, correcting the eyebrows, and shaping the mouth. Figure 3.2 presents an example of 3D facial heads created from 2D images.

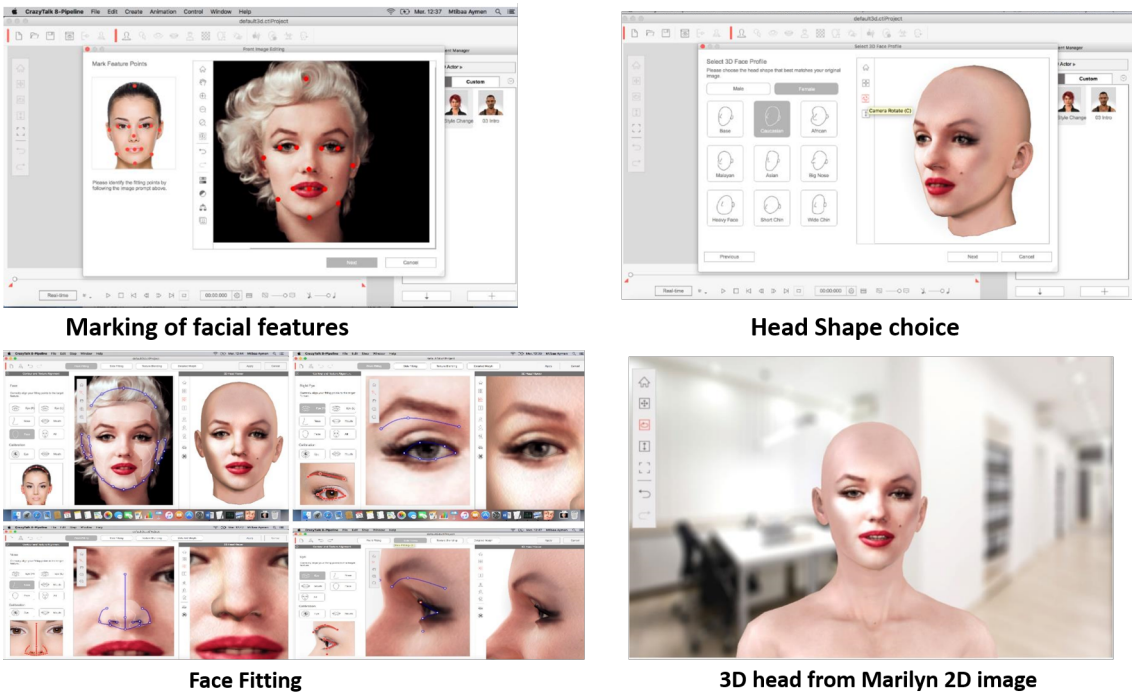


Figure 3.1: Steps for generating a 3D facial head from a 2D image using CrazyTalk.

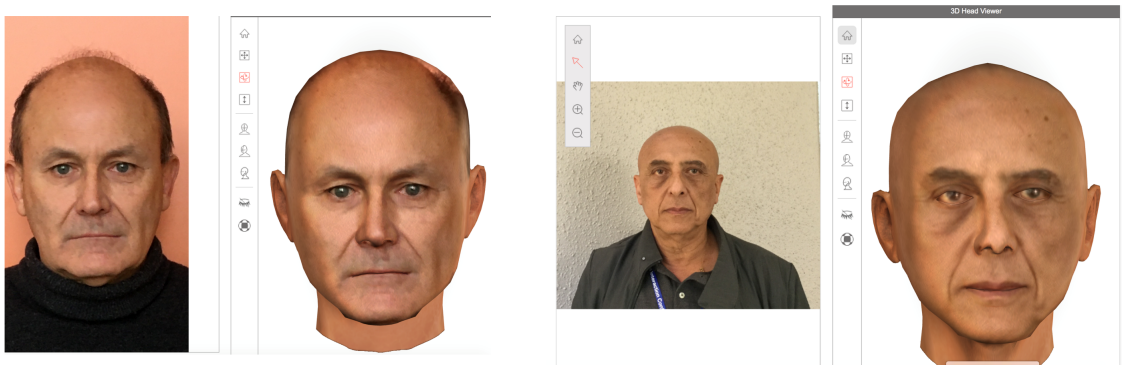


Figure 3.2: Examples of 3D facial heads generated from 2D images.

3.1.2 Animation of the 3D facial head

Facial recognition systems use anti-spoofing detectors that ask the user to perform specific movements (animations or expressions) in order to be authorized to access. Therefore, after generating the 3D head, we use iClone [3] tool to mimic these movements and produce a 3D head animated with facial expressions and movements, so that the 3D face models a real human face. This 3D head could also be animated and synchronized with voice samples to generate a 3D talking head. Figure 3.3 shows examples of facial expressions and animations that could be used to bypass liveness detectors.

Figure 3.3: Examples of facial expressions and movements (smile, blinking, raising eyebrows, rotating head) that could be animated with the 3D head to spoof the liveness detectors.

[3] https://www.reallusion.com/iclone/

3.2 Evaluation of SpeechXrays Audio-Visual Biometric System Against 3D Talking Head

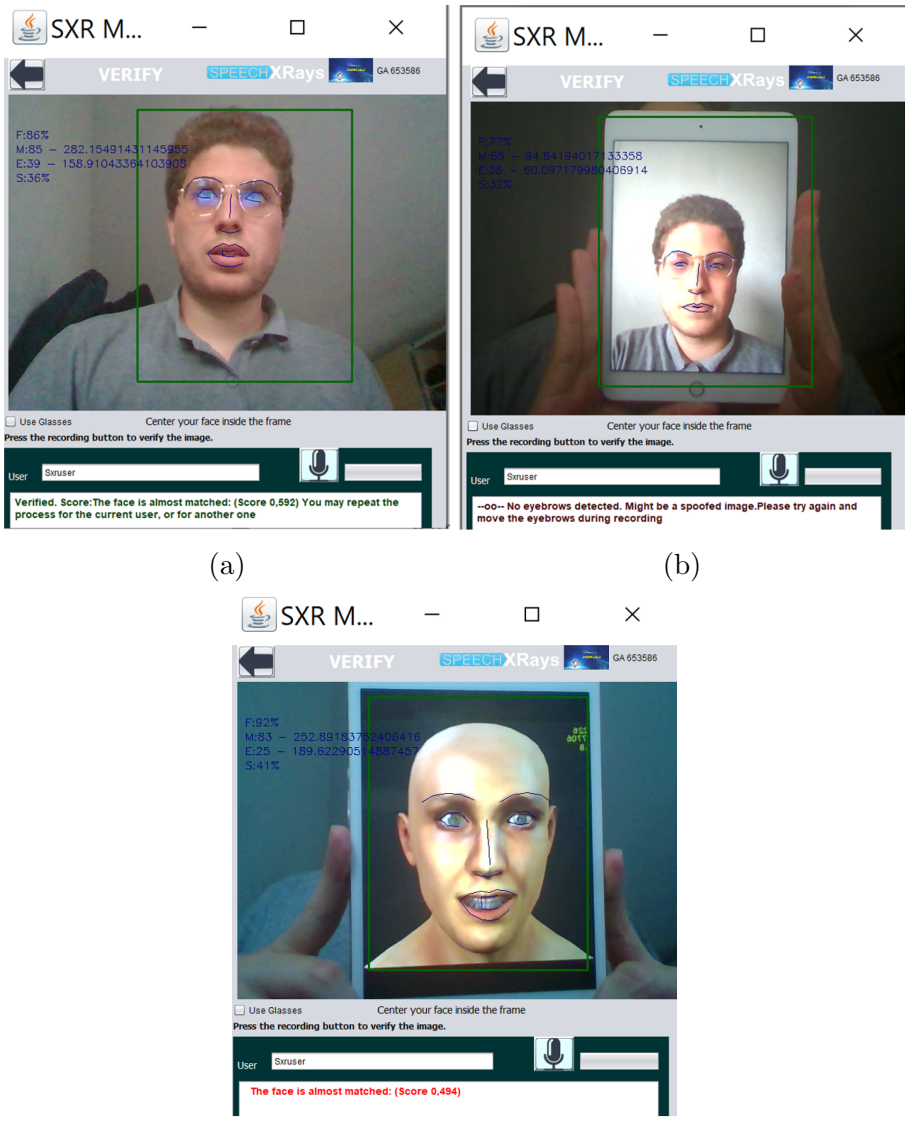
Using the methodologies described above for the creation of the 3D facial head, we evaluated the vulnerability of the audio-visual recognition system proposed during the H2020 SpeechXRays project against presentation attacks based on 3D talking head. For the SpeechXRays system, the user is asked to present his face and read a prompted sequence of digits to be authenticated.

We started the evaluation by analyzing the anti-spoofing strategies integrated into this audio-visual recognition system. This system incorporates certain liveness detectors into its authentication process. For face modality, it detects the physical presence of the target user by interpreting the movement of the lips and eyebrows, and by interpreting the colors and brightness of the image. For voice modality, the user is required to read the sequences of digits that appear during verification. Knowing this information, we get an idea of facial animations that we need to produce using the 3D facial head in order to bypass the anti-spoofing detectors. We focus our evaluation on bypassing the anti-spoofing detectors of the face recognition module using the 3D facial model. For the voice module, it is easier to reproduce the real target voice using impersonation, replay, speech synthesis, or voice conversion spoofing attacks. During our evaluation, we use the voice recordings of the target user to animate the 3D head.

For the evaluation, we started by enrolling a user in this audio-visual system. As a control, we first verified that the system can correctly authenticate the target user (Figure 3.4a). Next, before testing the system against 3D facial model attack, we evaluated its vulnerability against a fixed image of the target user. As shown in Figure 3.4b, the system resists such attacks and responds that no eyebrows movements are detected. Then, using an image of the target user, we created his 3D facial model animated with lips and eyebrows movements to bypass the liveness detectors. Finally, we recorded a video containing the animated 3D facial model and we play it to spoof the system. As shown in Figure 3.4c, the animated 3D facial model succeeds to bypass the liveness detectors and get a face verification score equal to 0.49 which is close to the score of a real user which is 0.59.

From this evaluation, we demonstrated the ability of an adversary to perform

a 3D facial reconstruction able to bypass the anti-spoofing detectors by using a 2D image of the target user. We have outlined that the fusion of speaker verification system with face modality is not enough to achieve secure authentication. In



(a) (b)

(c)

Figure 3.4: Authentication to the SpeechXRays audio-visual system; (a) target user authentication; (b) Impostor trial with fixed image of target user. (c) Spoofing of the audio-visual system using the animated 3D facial model of the target user created from his 2D image.

fact, there are several methods of 3D facial model reconstruction [Xu et al., 2016] and lip-synchronized facial animation generation [Taylor et al., 2017] that make it feasible to create a realistic 3D face model animated and synchronized with input audio. We believe that such methods pose a threat to the security and privacy of biometric systems and that one possible solution is the protection of speaker verification systems with privacy-preserving schemes.

3.3 Chapter Conclusions

In this chapter, we have presented a presentation attack based on 3D facial model created from a 2D image of the target user. This type of attack presents a real threat to audio-visual biometric systems. The fusion of the speaker verification system with facial recognition can be a solution to improve the biometric performance but it is not sufficient to guarantee the security and privacy aspects of the user biometric information. Therefore, in the next chapters, we propose as a solution, privacy-preserving schemes to improve the privacy and security of speaker verification systems.

Chapter 4

Privacy-Preserving Speaker Verification System Based on Cancelable Gaussian Mixture Models

Contents

This chapter presents a privacy-preserving speaker verification system based on Gaussian mixture model. This system includes two main stages: (i) transformation of speaker model into a binary representation (ii) the protection of the binary representation with a cancelable scheme named shuffling. The proposed system is evaluated according to the requirements of biometric information protection [ISO/IEC JTC1 SC27 Security Techniques, 2011] in terms of biometric performance, revocability, irreversibility, and unlinkability. Also, the robustness against different attack scenarios was analyzed.

The chapter is structured as follows. Section 4.1 gives a general description of the baseline speaker verification system based on GMM models. Section 4.2 presents the proposed cancelable GMM-based speaker verification system. A description of the architecture and the steps required to generate the cancelable biometric reference is provided. Evaluation of the proposed system is presented in Section 4.3. Finally, the chapter summary and conclusions are presented in section 4.4.

4.1 Baseline Speaker Verification System Based-GMM Models

One of the first successful approaches for speaker verification is the Gaussian mixture modeling [Reynolds et al., 2000]. Given the speech samples characterized by T-dimensional feature vectors $\mathbf{X}$ of Mel frequency cepstral coefficients (MFCC) $[\boldsymbol{x_1}, \boldsymbol{x_2}...\boldsymbol{x_T}]$, the user's features are presented by a GMM λ_s as follows:

$$P(\mathbf{x_t}|\lambda_s) = \sum_j \mathbf{w_j}\mathcal{N}(\mathbf{x_t}|\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j}) \tag{4.1}$$

$$\mathcal{N}(\boldsymbol{x_t}|\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma_j}|^{1/2}} exp\{-\frac{1}{2}(\boldsymbol{x_t} - \boldsymbol{\mu_j})'\boldsymbol{\Sigma_j}^{-1}(\boldsymbol{x_t} - \boldsymbol{\mu_j})\} \tag{4.2}$$

where w_j are the mixture weights and $\mathcal{N}(\mathbf{x_t}|\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})$ is the j^{th} multivariate Gaussian distribution with mean $\boldsymbol{\mu_j}$ and covariance $\boldsymbol{\Sigma_j}$. These parameters are trained using the enrollment voice samples using the expectation-minimization (EM) algorithm.

Although the speaker model can be extracted directly from the speaker enrollment data, it can also be generated from maximum a posteriori (MAP) adaptation using the universal background model (UBM) $\lambda_U = (\boldsymbol{w_i^U}, \boldsymbol{\mu_i^U}, \boldsymbol{\Sigma_i^U})$ where $i = 1, ..., M$ and M is the total number of Gaussian mixture components. The idea is to derive the speaker's model by updating the well-trained parameters in the UBM via adaptation. Given a UBM model λ_U and the enrollment features of the speaker $[\boldsymbol{x_1}, \boldsymbol{x_2}...\boldsymbol{x_T}]$, we first determine the probabilistic alignment of the features into the UBM mixture components by computing the posterior probabilities of the individual Gaussians in the UBM. For the i^{th} mixture component of the UBM, we compute:

$$P(i|\boldsymbol{x_t}) = \frac{\boldsymbol{w_i^U}\mathcal{N}(\boldsymbol{x_t}|\boldsymbol{\mu_i^U}, \boldsymbol{\Sigma_i^U})}{\sum_j \boldsymbol{w_j}^U\mathcal{N}(\boldsymbol{x_t}|\boldsymbol{\mu_j^U}, \boldsymbol{\Sigma_j^U})} \tag{4.3}$$

Then, we use the a posteriori probabilities $P(i|\boldsymbol{x_t})$ to compute the new mean, weights and variance parameters:

$$\boldsymbol{w_i}^{\text{`}} = \frac{1}{T}\sum_t P(i|\boldsymbol{x_t}) \tag{4.4}$$

$$\boldsymbol{\mu_i}^{\text{`}} = \frac{\sum_t P(i|\boldsymbol{x_t})\boldsymbol{x_t}}{\sum_t P(i|\boldsymbol{x_t})} \tag{4.5}$$

$$\boldsymbol{\Sigma_i}^{\text{`}} = \frac{\sum_t P(i|\boldsymbol{x_t})\boldsymbol{x_t}\boldsymbol{x_t^T}}{\sum_t P(i|\boldsymbol{x_t})} \tag{4.6}$$

Finally, we obtain the parameters of the adapted speaker model λ_s from the combination of the above parameters and the UBM parameters as follows:

$$\hat{\boldsymbol{w}}_{\boldsymbol{i}}^{\boldsymbol{s}} = \alpha_i \boldsymbol{w_i}^{\text{`}} + (1 - \alpha_i)\boldsymbol{w_i^U} \tag{4.7}$$

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{i}}^{\boldsymbol{s}} = \alpha_i \boldsymbol{\mu_i}^{\text{`}} + (1 - \alpha_i)\boldsymbol{\mu_i^U} \tag{4.8}$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{i}}^{\boldsymbol{s}} = \alpha_i \boldsymbol{\Sigma_i}^{\text{`}} + (1 - \alpha_i)[\boldsymbol{\Sigma_i^U} + \boldsymbol{\mu_i^U}\boldsymbol{\mu_i^{UT}}] - \hat{\boldsymbol{\mu}}_{\boldsymbol{i}}^{\boldsymbol{s}}\hat{\boldsymbol{\mu}}_{\boldsymbol{i}}^{\boldsymbol{sT}} \tag{4.9}$$

The adaptation coefficients α_i control the amount of contribution of the enrollment data relative to the UBM. During the verification phase, given the features of the probe sample $\boldsymbol{Y}$, we compute the likelihood ratio using the speaker model

λ_s and the UBM λ_U:

$$score = \frac{p(\boldsymbol{Y}|\lambda_s)}{p(\boldsymbol{Y}|\lambda_U)} \tag{4.10}$$

4.2 Cancelable GMM-Based Speaker Verification System

In this section, we present the proposed cancelable speaker verification system based on GMM. The idea was to combine binary representations of speakers' models that were originally developed for speaker modeling with cancelable schemes to achieve the privacy requirements. Therefore, the system is based on two steps: (i) transformation of the speaker's model into a binary representation, and (ii) the protection of the binary representation using a cancelable shuffling scheme.

4.2.1 Binary speaker representation

In [Anguera and Bonastre, 2010], a binarization approach to model the acoustic features with a binary vector was presented. This approach was exploited for speaker recognition task in [Bonastre et al., 2011] and for speaker diarization in [Anguera and Bonastre, 2011] and [Delgado et al., 2015]. For our system, we will use this binarization approach to develop a privacy-preserving speaker verification system.

The binarization method described in [Anguera and Bonastre, 2010] was based on Key background model to convert speaker utterances into binary vectors. In our work, a specific GMM for each speaker was used to extract the binary representation. Given a large set of speech data, first, a UBM is trained. Then, the speaker's GMM model is derived by adapting the enrollment utterances to the UBM. Next, a speaker's binary representation is defined as an N-dimensional binary vector, where N is the number of Gaussian Mixtures in the GMM model. Also, an accumulator vector initialized to 0 with the same length as the binary vector is defined. Each position in the binary vector will represent a Gaussian Mixture λ from the GMM model.

Given a speaker's utterance, the binary representation is extracted as shown

in step 1 of Figure 4.1. For each acoustic frame in the utterance, we compute the likelihood *lkld* given each of the Gaussians λ in the speaker's GMM model. Then, we select a percentage $\theta 1$ of Gaussians with the highest likelihood values. For the selected Gaussian, we increase by 1 the corresponding accumulator vector positions. When all frames have been processed, each position in the cumulative vector contains the relative importance of each Gaussian in modeling the utterance we have processed. The conversion of the accumulator vector into the binary representation is performed by setting the top $\theta 2$ percent positions in the accumulator vector with the highest values to one, and others to zero. $\theta 1$ and $\theta 2$ parameters should be set and optimized according to the biometric performance.

4.2.2 Cancelable speaker template

As shown in step 2 of Figure 4.1, after the binarization step, the speaker's binary representation is transformed using a shuffling scheme to generate the cancelable speaker template. The concept of shuffling scheme was introduced in [Kanade et al., 2012]. For each user, we associate a binary key K_{sh} of length L_{sh}. Then, the speaker's binary representation is divided into L_{sh} blocks each of the same length. To start the shuffling, these L_{sh} blocks are aligned with the L_{sh} bits of the shuffling key K_{sh}. In the next step, two distinct parts are created: the first part comprises all positions' blocks where the shuffling key bit value is one, and all the remaining blocks are taken in the second part. These two parts are concatenated to form the shuffled binary representation which is treated as the cancelable speaker template. The pseudo-code of the shuffling scheme is shown in Algorithm 1.

Based on this transformation, when two binary representations are transformed using the same shuffling key, the absolute positions of the blocks change but this change occurs in the same way for both of the representations. As a result, the distance between them keeps being the same. On the other hand, if they are shuffled using two different keys (impostor scenario), the result is a randomization of the representations, and the distance increases. In addition, this transformation makes it possible to generate different cancelable templates from the same binary representation by changing the shuffling key.

Figure 4.1 illustrates the steps required to generate the cancelable speaker

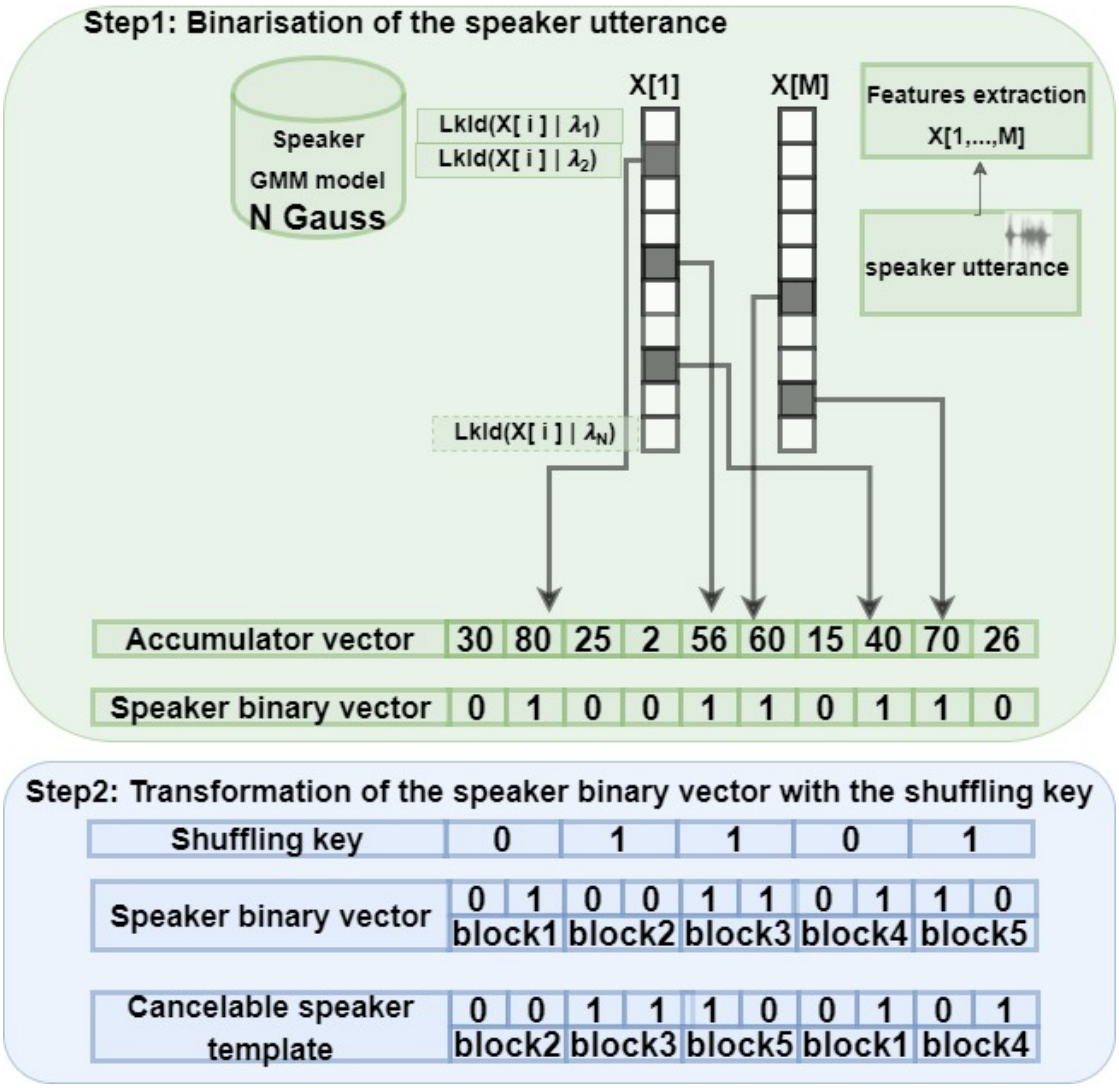


Figure 4.1: Steps required to generate the cancelable speaker template. Step1: binarization of the speaker's utterance. Step2: Transformation of the binary representation with the shuffling scheme.

template. First, we start by converting the provided utterance to a binary representation with the method described in subsection 4.2.1. Then, the binary representation is transformed with the user-specific shuffling key. The efficiency of this scheme is shown by its ability to affect only the alignment not the values of the binary vector-bits. This is important because each bit-value in the binary representation is the projection of the acoustic location of each acoustic frame from

Algorithm 1: Shuffling scheme pseudo-code

shuffling $(data, K_{sh})$;
Input: $data$: Binary speaker representation, K_{sh}: shuffling key
Output: cancelable template
Initialization
part 1 = [ ] Define empty vector
part 2 = [ ] Define empty vector
L_{sh} = Length of the shuffling key (K_{sh})
L_{data} = Length of the binary representation $(data)$
$block_{size}$= L_{data} / L_{sh}
$j \leftarrow 1$
for $i = 1 : L_{sh}$ **do**
  if $K_{sh}(i) = 1$ **then**
    part 1 $\leftarrow [part1, data(\text{ j ; j + block}_{size} - 1)]$
  else
    part 2 $\leftarrow [part2, data(\text{ j ; j + block}_{size} - 1)]$
  end
  $j \leftarrow j + block_{size}$
end
cancelable template $\leftarrow$ *concatenate* $[part1\ ;\ part2]$

the feature space into the space of GMM Gaussian. Besides, the shuffled binary vector, which is treated as the cancelable template, is the result of combining the biometric sample and the shuffling key. Therefore, once it is leaked, it can be revoked and a new template can be generated by changing the shuffling key.

4.2.3 System architecture and protocol of the cancelable speaker verification system based on GMM

Figure 4.2 illustrates the architecture of the proposed cancelable speaker verification system based on GMM. As input, we assume that the access server already has the UBM trained on publicly available data and that the shuffling key of the user is stored in his token. A unique shuffling key is assigned to each user during enrolment and he/she has to provide that same key during each subsequent verification.

During the enrollment phase, the system sends the UBM to the client-side

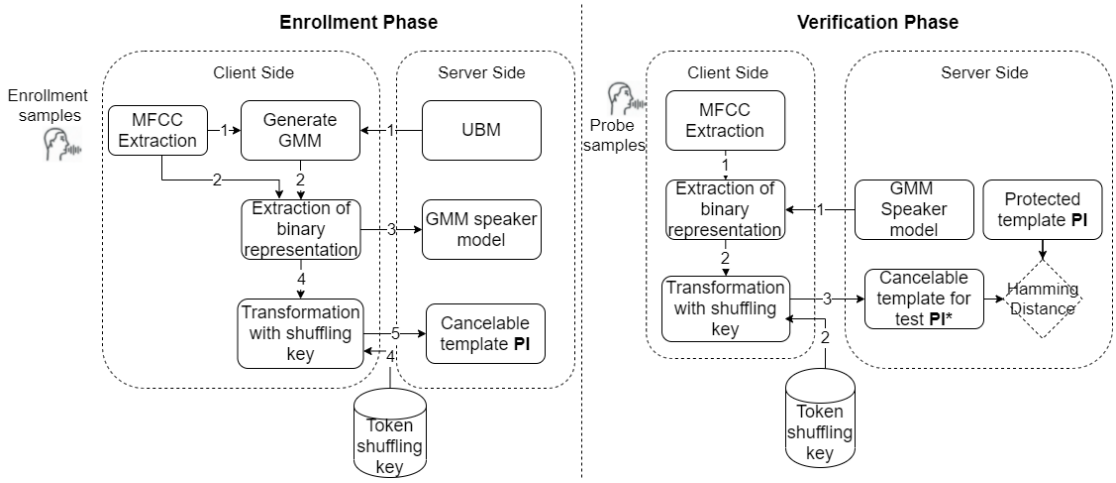


Figure 4.2: Architecture of the privacy-preserving speaker verification system based on cancelable GMM.

that performs the adaptation with the enrollment features to generate the adapted model GMM. Using the user's GMM, the client-side converts the samples of enrollment into a binary vector representation. Then, the binary representation is transformed using the user's shuffling key received from the token to generate the enrollment cancelable template named pseudonymous identifier *PI*. In the end, the client-side sends the cancelable template and the user's GMM to the server. After execution of the enrollment phase with all users, the access server has the protected templates, along with the GMMs.

During the verification phase, the server sends the GMM of the claimed identity to the client-side to extract the binary representation from the probe samples. Then, the token sends the shuffling key to the client-side that transforms the binary representation and generates the probe cancelable template *PI**. The *PI** is transferred to the server that measures the Hamming distance between the stored *PI* and the *PI** to decide based on a predefined threshold the outcome of the verification. To compare between two cancelable templates *PI* and *PI**, a dissimilarity score s is obtained by computing their Hamming distance as follows:

$$s(PI, PI^*) = 1 - \frac{\sum_{i=1}^{N} (PI[i] \wedge PI^*[i])}{N} \tag{4.11}$$

where $\wedge$ is the operator of AND logic between any two bits.

Based on this protocol, during the enrollment and verification, the server never has access to the voice samples provided by the user. The client-side sends only the protected templates to the server sides, and the biometric comparison is performed with the cancelable templates on the transformed domain. A shortcoming of the above protocol is that the server has the GMMs in plaintext. One possible solution is to store the speaker's GMM on the client-side or to encrypt the GMM before sending it to the server.

4.3 Experimental Evaluation and Results

In this section, the proposed privacy-preserving speaker verification system based on GMM is evaluated according to the privacy requirements described in the standard for biometric information protection [ISO/IEC JTC1 SC27 Security Techniques, 2011]. First, biometric performance evaluation of speaker verification systems based on the baseline (unprotected) GMM and the cancelable templates will be reported. Then, the evaluations of the revocability, unlinkability, and irreversibility are provided. Furthermore, a security analysis of the cancelable speaker verification system based against different attack scenarios is reported.

4.3.1 Databases

The experiments are conducted using TIMIT database [Keating et al., 1994] for tuning and parameterizing the speaker binary representation, and the RSR2015 text-dependent database [Larcher et al., 2014] to evaluate the cancelable system.

TIMIT database contains a total of 6300 sentences, ten sentences spoken by each of the 630 speakers (438 males and 192 females) of eight major dialects of American English. This database was used to tune the speaker binary representation.

The RSR2015 database comprises speech recorded from 300 speakers, including 143 females and 157 males. For our evaluation, part1 of RSR2015 is used. This part focuses on a text-dependent speaker verification task where each speaker pronounces 30 fixed sentences in nine sessions. The duration of each sentence

varies between 2 and 3 seconds. The comparison protocol described in [Larcher et al., 2014] is followed. From the nine sessions of each speaker, three sessions are used for the enrollment while the rest of the sessions are used for the test.
RSR2015 database provides four types of trials depending on whether the test utterance is spoken by the target user or not and that the spoken utterance is the correct passphrase or not:
Target-correct (tar-c): where the target speaker pronounces the expected pass-phrase.
Target-wrong (tar-w): where the target speaker pronounces a wrong pass-phrase (a phrase that is different from the enrollment one).
Impostor-correct (imp-c): where a non-target speaker pronounces the expected pass-phrase.
Impostor-wrong (imp-w): where a non-target speaker pronounces a wrong pass-phrase (a phrase that is different from the enrollment one).
Target correct trials are considered as target trials, while the others are considered as non-target trials. The impostor-correct trials are more challenging, as the non-target user pronounces the expected passphrase that is used to enroll the target speaker.

4.3.2 Experimental setting

The feature extraction component is common for the baseline GMM-UBM system and the proposed cancelable system. The feature vector is composed of 20 MFCC coefficients with their first and second derivative coefficients and the log energy leading to a 63-dimensional feature vector. The MSR Identity Toolbox [Sadjadi et al., 2013] was used to extract the features.

As described in section 4.2, for both the baseline GMM-UBM and the cancelable system, we need to train a GMM model for each speaker. For this, UBM gender-dependent models are trained with 1024 Gaussians using the background partition of RSR2015 database. Then, the speaker GMM model is trained by adapting the enrollment utterances to the UBM using the MAP criterion. As described in the protocol of RSR2015 part1, three utterances $enrollspeech_g$ selected from the nine sessions are used to train the speaker's GMM model during the

enrollment.

For the cancelable system, during the enrollment phase, the selected enrollment utterances $enrollspeech_g$ used to train the specific speaker GMM_g model are employed to extract three binary vectors using the steps described in Figure 4.1. Then, from the three binary vectors, the speaker binary representation of length 1024 bits is extracted by considering the significant bits (bits in the binary vectors which are less likely to change). The transformation of the speaker binary representation with the user shuffling key key_g generates the enrollment cancelable template. For our evaluation, we use shuffling keys with length equal to the speaker binary representation $L_{data} = L_{sh} = 1024$.

During the verification phase, the user presents the probe voice samples and the shuffling key. The key could be the same as the enrollment key in the case of a genuine probe or it could be a random key in the case of an impostor probe. For genuine comparison, the probe cancelable template of the target user is extracted from his/her GMM_g by providing the target probe voice samples and the target user shuffling key key_g. For impostor trials, a non-target user I will try to extract a probe cancelable template from the target user GMM_g model by providing his/her probe voice samples $probe - speech_I$, and his/her shuffling key key_I.

4.3.3 Binary speaker representation analysis

In order to extract speaker binary representations that discriminate between speakers' characteristics, the parameters $\theta 1$ and $\theta 2$ need to be tuned according to the biometric performance. In this evaluation, we search the optimum parameters $\theta 1$ and $\theta 2$ that minimizes the equal error rate (EER). Using TIMIT database, for each speaker, we extract the different possibilities of binary vectors according to parameters $\theta 1$ and $\theta 2$. Figure 4.3 shows the EER distribution on TIMIT database for SV system based on binary speaker vectors according to $\theta 2$ for a fixed value of $\theta 1$ equal to 2%. As shown, for $\theta 2 < 20\%$ binary vectors cannot discriminate between speakers, because the most selected positions coincide with Gaussians that model noisy acoustic frames. Also, for $\theta 2 > 40\%$ the discriminability power of the speaker binary vector degrades. We observe that the optimum value for $\theta 2$ is 30%. Running a similar experiment for $\theta 1$, we observe that $\theta 1 = 2\%$ minimize the EER.

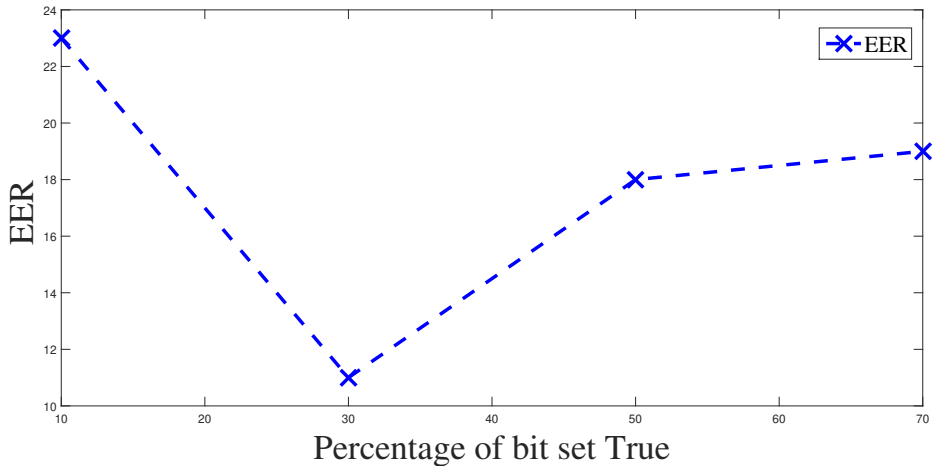


Figure 4.3: EER distribution for speaker verification system based on binary representation for the speaker according to parameter $\theta 2$ on the development TIMIT database. The $\theta 2$ parameter tuned on TIMIT database will be used during the evaluation on the RSR2015 database.

For the rest of the work, we use $\theta 1$ equal to 2% and $\theta 2$ equal to 30% for the extraction of the binary representations on the RSR2015 database.

4.3.4 Biometric performance evaluation of the cancelable GMM system

One of the main requirements for cancelable biometrics is the fact that the protection of biometric information should not degrade the biometric performance compared to the baseline system (unprotected system). Therefore, for objective comparison, the biometric performance of the baseline GMM-UBM and the proposed cancelable system are reported.

In this evaluation, we report the performance of the cancelable system in the legitimate scenario. In this scenario, the target user employs his probe biometric sample with his shuffling key to be authenticated, and the non-target user will use his probe biometric sample with a random shuffling key to impersonate the target user. The system performance is reported in terms of Equal Error Rate (EER). The EER is the rate at which the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) are equal. We have also used the targets and no-targets score

Table 4.1: Biometric performance of the speaker verification systems based on the baseline GMM, binary templates, and cancelable templates on the RSR2015 female evaluation subset for the impostor correct and impostor wrong trials in terms of EER (%).

Performance EER%	**Baseline GMM**	**Binary template**	**Cancelable GMM**
tar-c/imp-c	1.98	10.25	0.01
tar-c/imp-w	0.43	2.62	0.01

Table 4.2: Biometric performance of the speaker verification systems based on the baseline GMM, binary templates, and cancelable templates on the RSR2015 male evaluation subset for the impostor correct and impostor wrong trials in terms of EER (%).

Performance EER%	**Baseline GMM**	**Binary template**	**Cancelable GMM**
tar-c/imp-c	3.5	16.05	1.32
tar-c/imp-w	0.9	7.31	2.18

distributions along with the ROC curves to evaluate the matching performance.

Regarding the biometric performance of the baseline GMM speaker verification system reported in Tables 4.1 and 4.2, we observe that better biometric performance is obtained with target-correct/impostor-wrong trials than with the target-correct/impostor-correct trials. This was expected since the impostor-correct trials are more challenging, as the non-target user pronounces the expected passphrase used by the target user to authenticate.

For the speaker verification system based on binary representations, the EER degrades compared to the baseline system due to the loss of biometric information. For example, on the female subset, for target-correct/impostor wrong trails, the EER increases from 0.43% using baseline GMM to 2.62% using binary templates.

For the cancelable GMM system, the biometric performance obtained in the female and male evaluation subset of RSR2015 are respectively reported in Tables 4.1 and 4.2. The proposed system improves the biometric performance compared to the baseline system. A clear improvement in terms of EER for both trials impostor correct and impostor wrong is reported. As an example, on the female subset, for target-correct/impostor-correct trails, the EER of the baseline system is 1.98%

which reduces to 0.01% when shuffling scheme is applied to the binary templates.

The biometric performance improvement is related to the overlap between the target and non-target distributions. The smaller the overlap between the two distributions, the better the system performs. Through the distributions in Figure 4.4, it can be shown that the shuffling scheme preserves the target Hamming distances and increases the non-target Hamming distances. When applying the shuffling scheme, the mean of the target distribution is preserved exactly just like in the binary templates level before performing the shuffling transformation. Contrarily, the mean of the non-target distribution is augmented when the shuffling scheme is applied and the distribution is right-shifted. This reduces the overlap between target and non-target distributions which improves the discrimination capacity of the system and thereby leads to a better verification performance.

We report in Figure 4.5 the ROC curves obtained for speaker verification systems based on the baseline GMM, binary templates, and the cancelable templates on the female evaluation subset of RSR2015 for the target-correct/impostor-wrong

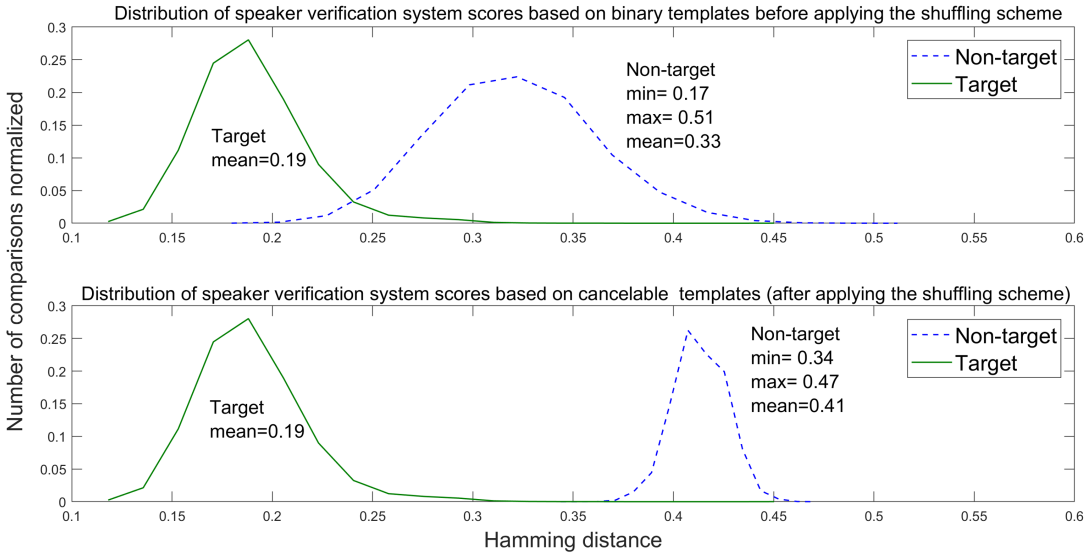


Figure 4.4: Distribution of speaker verification systems scores based on the binary templates (before applying the shuffling) and cancelable templates (after applying the shuffling). The distributions are reported for target-correct and impostor-correct trials on the female evaluation subset of part1 RSR2015 database in the legitimate scenario.

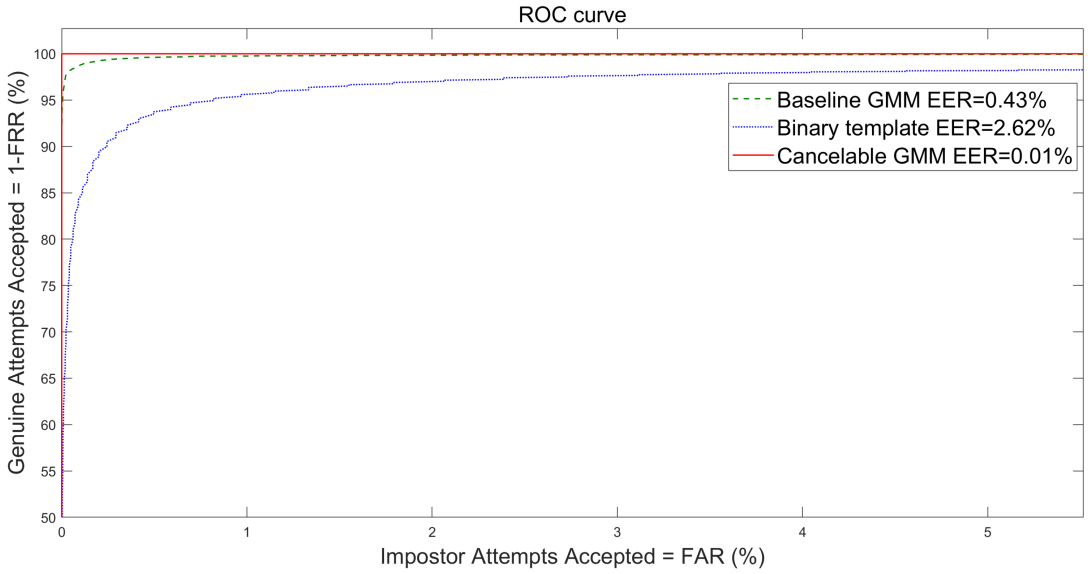


Figure 4.5: ROC curves for speaker verification systems based on the baseline GMM, binary speaker representation and the proposed cancelable templates on the female evaluation subset of part1 RSR2015 database.

trials. The cancelable system reaches better results with an EER = 0.01% compared to the baseline GMM system with EER = 0.43%

4.3.5 Revocability analysis

For the cancelable biometric system, the protected biometric template should be able to be revoked and renewed in case it is compromised. Revocability is evaluated by calculating the pseudo-impostor scores. The pseudo-impostor score is the comparison of a cancelable template of a particular user with other cancelable templates of the same user generated from the same biometric sample and transformed with different shuffling keys. For this, we transformed a speaker's binary template with 300 000 randomly generated shuffling keys. Then, the first shuffled template is compared with the remaining cancelable templates to compute the pseudo-impostor scores. This process is repeated with 30 different users. As shown in Figure 4.6, the distribution of the pseudo-impostor scores resembles the non-target distribution which means that the generated shuffling templates are indistinguishable from each other, although they are generated from the same

speaker's binary template. As a result, in case of compromise, a cancellation is possible and a new cancelable template can be generated by changing the shuffling key.

For the protection of the binary representations using the shuffling scheme transformation, the maximum number of the cancelable templates or Pseudonymous Identifier PI that can be generated from the same biometric sample is given by the number of possible permutations. Moreover, because the decision in the proposed system is based on a threshold comparison, we should not account for templates falling in the same neighborhood. We estimate the maximum number of templates using the Hamming-packing bound [MacWilliams and Sloane, 1977].

We assume that the target speaker template is the center of a sphere with a radius of r, known as a Hamming sphere. r represents the maximum number of non-matching bits obtained when comparing two templates belonging to the same speaker. r is equal to $(t \times l)$ where t is the EER threshold of the cancelable system and l is the length of cancelable template. Then, the possible templates, that their distance compared to the speaker template are less than the radius r (meaning they are within the sphere) are not taken into account. Using the EER threshold $t = 0.37$ of the cancelable system, for speaker template of length $l = 1024$ and shuffling key of $L_{sh} = 1024$, we get almost 2^{50} possible cancelable template PI for each user as given in Eq. 4.12.

$$PI = \frac{Number\ of\ possible\ permutation}{Volume\ of\ Hamming\ spheres} = \frac{1024!}{(512!)(512!)\sum_{k=0}^{(t\times 512)/2}\binom{512}{2k}} \approx 2^{50} \tag{4.12}$$

4.3.6 Unlinkability analysis

As defined in [ISO/IEC JTC1 SC27 Security Techniques, 2011] the unlinkability is "a propriety of two or more biometric references that they cannot be linked to each other or to the subject(s) from which they were derived". The goal of this evaluation is to determine if there exists some methods to decide if two protected templates T1 and T2 enrolled in different applications are generated from the

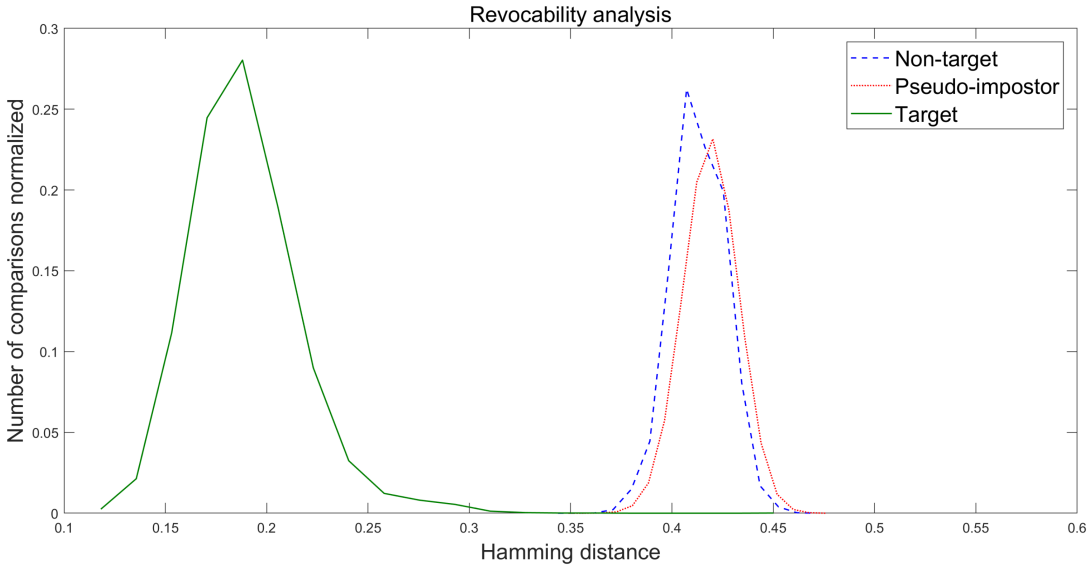


Figure 4.6: Revocability analysis: Distribution of target, non-target, and pseudo-impostor scores for cancelable GMM system on the female evaluation subset of part1 RSR2015 database.

same biometric sample or not. For this, we use the framework defined in [Gomez-Barrero et al., 2017] to evaluate the unlikability of the proposed cancelable speaker verification system. Two types of score distributions will be analyzed for the assessment of the unlinkability provided by the protected templates:

Mated instances H_m*:* scores computed from cancelable templates extracted from different samples of the same subject using different shuffling keys. It represents the probabilities $p(s|H_m)$, where s is the score between two templates.

Non-mated instances H_{nm}*:* scores computed from cancelable templates generated from samples of different subjects using different shuffling keys. It represents the probabilities $p(s|H_n m)$.

$D_{\leftrightarrow}^{sys} \in [0,1]$ was defined in [Gomez-Barrero et al., 2017] to have an estimation of the global linkability of the system:

$$D_{\leftrightarrow}^{sys} = \int_{S_{min}}^{S_{max}} D_{\leftrightarrow}(s) p(s|H_m)\, \mathrm{d}s \tag{4.13}$$

where $D_{\leftrightarrow}(s) \in [0,1]$ gives an estimation of the linkability of a system for a specific score and $[S_{min}, S_{max}]$ is the whole score range. If a system has $D_{\leftrightarrow}^{sys} = 1$, where

both score distributions (mated and non-mated) have no overlap, it means that the system is fully linkable. If a system has $D_{\leftrightarrow}^{sys} = 0$, where both score distributions (mated and non-mated) are overlapped, it means that the system is fully unlinkable. As shown in Figure 4.7, the distribution of mated and non-mated scores for the cancelable GMM system overlap with $D_{\leftrightarrow}^{sys}$ equal to 0.1, which makes the system fully unlinkable.

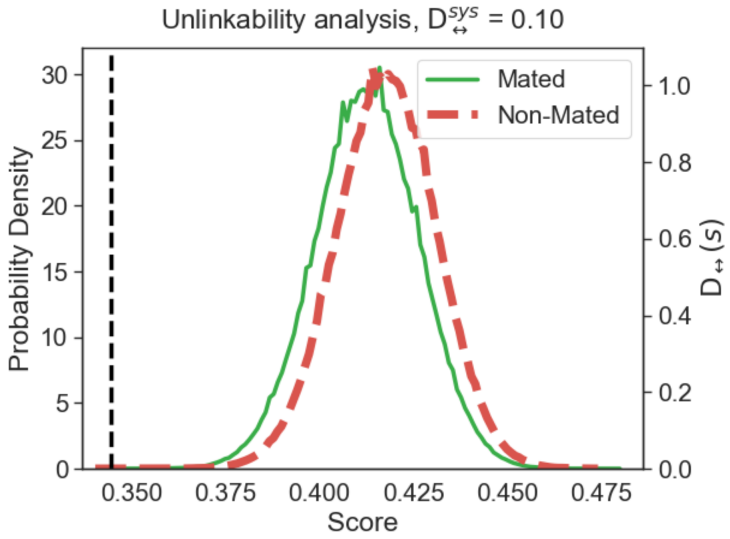


Figure 4.7: Unlinkability analysis: Distribution of Mated and Non-Mated scores for the cancelable GMM system on the female evaluation subset of part1 RSR2015 database.

4.3.7 Irreversibility analysis

The irreversibility refers to the security of the biometric feature from which the cancelable template was generated [ISO/IEC JTC1 SC27 Security Techniques, 2011]. For the proposed system, the reversibility analysis depends on whether the attacker has information about the shuffling key or not.

Given only the shuffling key, the attacker could not reconstruct the original binary representation, since the shuffling key does not provide information about the values of the binary vector. Without having information about the shuffling key and prior knowledge about the distribution of the non-shuffled binary vectors,

it is computationally not feasible to revert to the original binary representation as the number of permutations to be tested is too big. In the proposed system, if the adversary wants to guess the correct value of binary vector with a length of 1024 and knowing that the number of bits equal to 1 is 30% of bits, the guessing complexity is equal to 2^{395} the number of possible permutation, given by Eq. 4.14 as follows:

$$Number\ of\ possible\ permutation = \frac{1024!}{(1024 \times 0.3)!(1024 \times 0.7)!} \approx 2^{1018} \quad (4.14)$$

In case the attacker has stolen the shuffling key and the cancelable template, the reconstruction of the original binary representation is feasible. However, due to the binarization process, it is difficult to recover the original features.

4.3.8 Security analysis

The proposed system involves two factors: biometric and shuffling key. In real-world applications, it is mandatory to evaluate the system in the following scenarios:

Stolen biometric scenario: In this scenario, an attacker uses the biometric sample of the target user and transforms it with a random shuffling key to pretend as a target user.

Stolen shuffling key scenario: In this scenario, the attacker has the shuffling key of the target user and tries to access the system by presenting his/her biometric sample and the shuffling key of the target user.

4.3.8.1 Stolen biometric attack

To evaluate the proposed cancelable system against stolen biometric attacks, we compute the false acceptance rate (FAR), when the EER threshold of the cancelable system is considered as the decision threshold. We suppose that the biometrics data for all the speakers in the RSR2015 database are compromised. In such a case, an adversary provides the stolen biometric with a random shuffling key to gain access as the target user. In Figure 4.8, we report the FAR curve obtained in such an attack scenario. As shown, at the EER threshold=0.37, the FAR is equal

to 0. Thus, the system is robust to stolen biometric attacks.

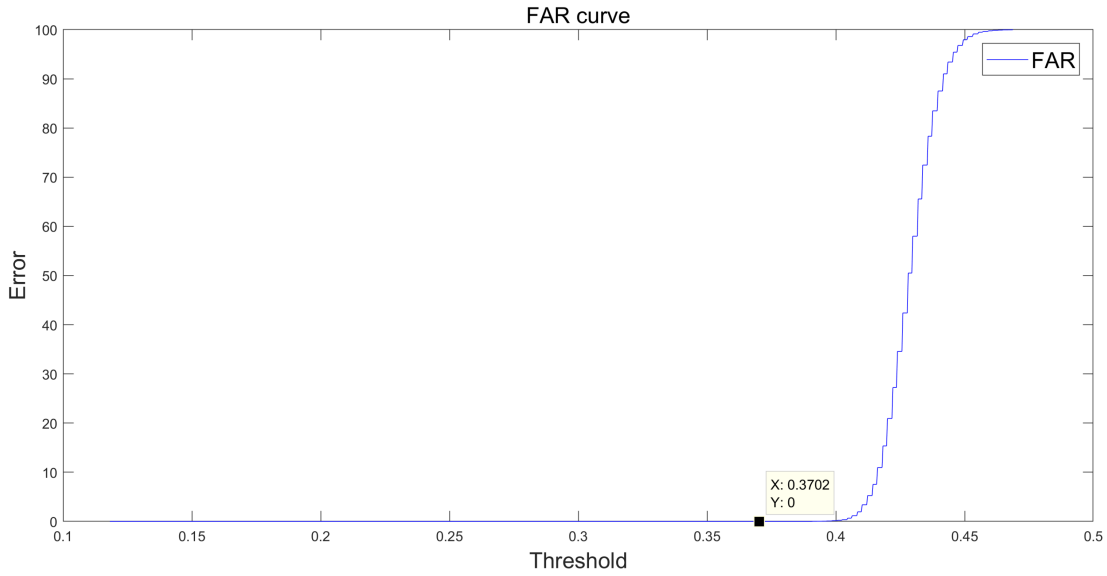


Figure 4.8: Stolen biometric analysis: FAR curve of the cancelable GMM system in the stolen biometric attack scenario.

4.3.8.2 Stolen shuffling key attack

Table 4.3: FAR of stolen key attacks according to the the cancelable system performance in terms of FAR and FRR in the legitimate scenario.

Legitimate scenario	**FAR %**	0	0	0	0	0	0	0
	FRR %	0.32	0.79	1	1.54	2	2.6	3
FAR stolen key scenario%		21.4	14.4	10.9	6.48	3.55	2.6	1.79

In this scenario, the attacker uses his/her biometric sample and the target user's shuffling key to gain unauthorized access. In Table 4.3, we report the false acceptance rate obtained for this attack according to the biometric performance of the cancelable system in terms of FAR and FRR in the legitimate scenario. The FAR in the stolen-key scenario depends on the FRR fixed for the cancelable system in the legitimate scenario. In fact, by increasing the FRR of the cancelable system in the legitimate scenario, we improve the FAR for the stolen key scenario. As shown in Figure 4.9, when we reduce the threshold of the verification decision,

for the legitimate scenario, we still get FAR=0 but the FRR increases. However, for stolen key scenario, the system becomes more robust since the FAR for stolen key attack decreases. As reported in Table 4.3, at the threshold corresponding to FRR=1.54% and FAR=0 in the legitimate scenario, the FAR=6.48% in the stolen key scenario. When we choose the threshold corresponding to FRR=3% and FAR=0 in the legitimate scenario, we improve the robustness of the cancelable system to stolen key attack and a FAR=1.79% is obtained.

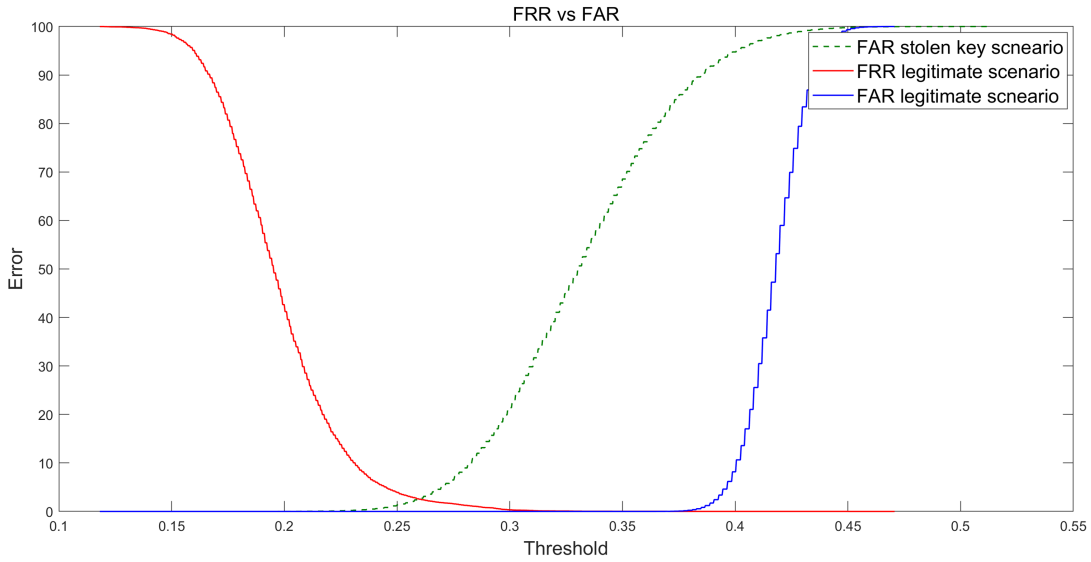


Figure 4.9: Stolen shuffling key analysis: FAR and FRR curves of the cancelable GMM system in the legitimate and the stolen key scenarios.

4.3.8.3 Brute force attack

A brute force attack consists of an adversary trying to guess the correct cancelable template to access as the target user. In the proposed system, the verification decision is based on Hamming distance comparison. The dissimilarity score is computed based on the number of matches between the enrollment and probe cancelable templates. If the dissimilarity score is less than the threshold, the test will be deemed as a legitimate user. For the proposed system, the dimension of the cancelable template is 1024-bits, and the threshold at EER=0.01% is 0.37. Therefore, if the adversary wants to guess the correct value of the cancelable

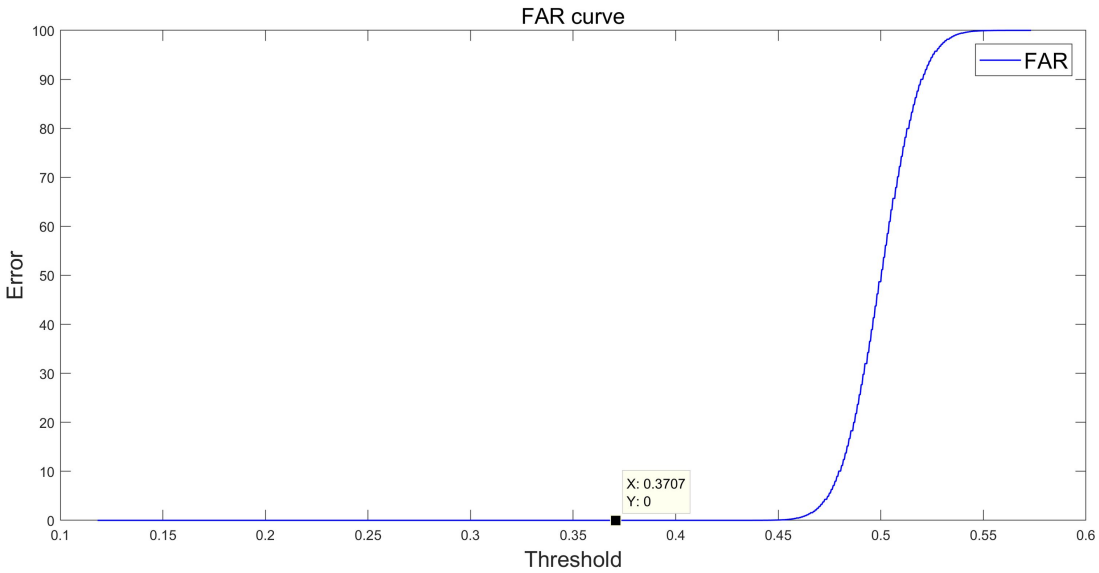


Figure 4.10: Brute force attack: FAR curve of the cancelable GMM system in the brute force attack scenario. FAR=0 at the EER threshold.

template, the guessing complexity is $2^{1024*(1-0.37)}$ attempts.

For the evaluation of the brute force attack, we attacked the cancelable templates of 30 user of part1 RSR2015 with 200,000 synthesized templates. As shown in Figure 4.10 the FAR = 0 for this attack at the EER threshold.

4.4 Chapter Summary and Conclusions

In this chapter, we have proposed a privacy-preserving speaker verification system based on GMM. Cancelable templates are generated by first modeling the speaker with a binary template. Then, the binary template is protected with the shuffling scheme. Experimental results show that the speaker verification system based on cancelable templates improves the biometric performance compared to the baseline GMM system. The system achieves the privacy requirements while maintaining the biometric performance and an EER = 0.01% was reported. In addition, the proposed system satisfies the requirements of biometric information protection described in the ISO/IEC 24 745. The transformation of the speaker's binary template with the shuffling scheme makes it possible to generate from the same biometric sample different versions of cancelable templates that cannot be linked to the user. These properties ensure the privacy of the user when he is enrolled in different applications using the same biometric sample (prevents cross-matching), and in case the user's cancelable template is compromised, it will be revoked and renewed.

Chapter 5

Privacy-Preserving Speaker Verification System Based on Cancelable i-Vectors

Contents

This chapter presents a privacy-preserving speaker verification system based on a cancelable i-vector. The cancelable scheme includes two steps, (i) i-vector binarization, and (ii) the protection of the binary i-vector with the shuffling scheme. The proposed system performs speaker verification without revealing the speaker's voice information to the access server, either during enrollment or during the verification phase. Unlike the cancelable GMM system proposed in chapter 4, the system based on cancelable i-vector doesn't require storing the speaker's GMM in plaintext. Privacy evaluation of this system according to the standard of biometric information protection (ISO/IEC 24745) shows that the proposed cancelable i-vector system achieves the revocability, unlinkability, irreversibility requirements, and improves biometric performance compared to the unprotected system. Moreover, security analysis was performed based on the evaluation methodology described in [Rosenberger, 2018]. Additionally, we demonstrate that the proposed cancelable scheme can also operate to protect deep neural network speaker embeddings such as x-vectors.

This chapter is structured as follows. Section 5.1 gives a general description of the baseline speaker verification system based on i-vectors. In Section 5.2, we present the proposed speaker verification system based on cancelable i-vector. A description of the architecture and the steps required to generate the cancelable template is provided. Evaluation of the proposed system is presented in Section 5.3. Finally, the chapter summary and conclusions are presented in section 5.4.

5.1 Baseline Speaker Verification Based on non-Protected i-Vectors

The i-vector system proposed by [Dehak et al., 2010] provides a way to generate a low dimensional fixed-length representation of a speech utterance that preserves speaker-specific information. This technique was inspired by the Joint Factor Analysis framework presented in [Kenny et al., 2008]. The i-vector system maps a sequence of features such as MFCC obtained from a speech utterance to a fixed-length low dimensional vector. A Universal Background Model is used to collect Baum-Welch statistics from the speech utterance. Then, the speaker-and channel-

dependent GMM-supervector $\mathbf{M}$ is constructed by appending together the first-order statistics for each mixture component that can be represented via a single total variability subspace as follows:

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w} \tag{5.1}$$

where $\boldsymbol{m}$ is the speaker- and channel-independent supervector extracted from the UBM, $\boldsymbol{T}$ is a low-rank matrix named total variability matrix spanning the subspace with speaker-specific information variability, and $\boldsymbol{w}$ is a standard normal distributed vector. The posterior mean of $\boldsymbol{w}$ is the corresponding i-vector.

The i-vector comprises both speaker and channel variability. Therefore, channel compensation or channel modeling techniques usually follow the i-vector extraction process as the Linear Discriminant Analysis (LDA) or Within-Class Covariance Normalization (WCCN) [Kanagasundaram et al., 2011]. For the biometric comparison, the cosine scoring is used to compare the target speaker i-vector $\boldsymbol{w}_{target}$ and the probe i-vector $\boldsymbol{w}_{test}$:

$$score(\boldsymbol{w}_{target}, \boldsymbol{w}_{test}) = \frac{\langle \boldsymbol{w}_{target}, \boldsymbol{w}_{test} \rangle}{\|\boldsymbol{w}_{target}\| \, \|\boldsymbol{w}_{test}\|} \tag{5.2}$$

Also, the probabilistic linear discriminant analysis (PLDA) [Kenny, 2010], [Garcia-Romero and Espy-Wilson, 2011] was introduced as back-end scoring. PLDA has the advantage of producing well-calibrated likelihood ratios without requiring score normalization when training and evaluation data are drawn from the same domain. For the proposed cancelable i-vector, we address the protection of i-vector system using cosine distance as back-end scoring.

5.2 Cancelable Speaker Verification System Based on i-Vectors

In this section, we describe the cancelable speaker verification system based on the binarization of i-vector and its transformation with the shuffling scheme.

5.2.1 Binary i-vector representation

The goal behind the binarization step is to hide the original i-vector. In order to extract a binary representation from the speaker's i-vector, thresholding method was applied. The use of the mean or median of the i-vectors as threshold gives close results on binary i-vectors since i-vectors distribution is close to the normal distribution. For the proposed system, the median is used to be sure that independently of the speaker, each binary i-vector contains an equal number of ones and zeros. This is useful in the revocability and irreversibility analysis. Given a speaker's i-vector, the elements having a higher value than the median are converted to one, while the remaining are converted to zero. From an i-vector $\mathbf{X}$ of dimension N, $\mathbf{X} = (x_1, ..., x_N)$, we obtain a binary vector $\mathbf{X}_{bin} = (b_1, ..., b_N)$ by comparing each component to the median value of the i-vector.

$$b_i = \begin{cases} 0, & \text{if } x_i \leq median(\boldsymbol{X}) \\ 1, & \text{otherwise} \end{cases} \quad \text{for } i \text{ in } \ (1, ..., N) \tag{5.3}$$

5.2.2 Cancelable i-vector

After i-vector binarization, the binary i-vector is transformed with the shuffling scheme [Kanade et al., 2012] described in Chapter 4, section 4.2.2, subsection 4.2.2, to generate the cancelable i-vector. The cancelable i-vector template is the result of combining the biometric sample (binary i-vector) and the shuffling key. Therefore, once the protected i-vector is leaked, it can be revoked and a new template can be generated by changing the shuffling key.

Figure 5.1 illustrates the architecture of the proposed cancelable i-vector. According to the ISO/IEC 24745, the system falls under the category Model G. This model employs data separation through distributed storage of data elements. We propose the following protocol for the proposed cancelable system. As input, we assume that the server already has the total variability matrix T with the UBM and the shuffling key of the user is stored in the token.

During the enrollment phase, the user provides the enrollment voice samples to the client-side that extracts the $MFCC$ features and generates the binary i-vector using the total variability matrix T and UBM received from the server. Then,

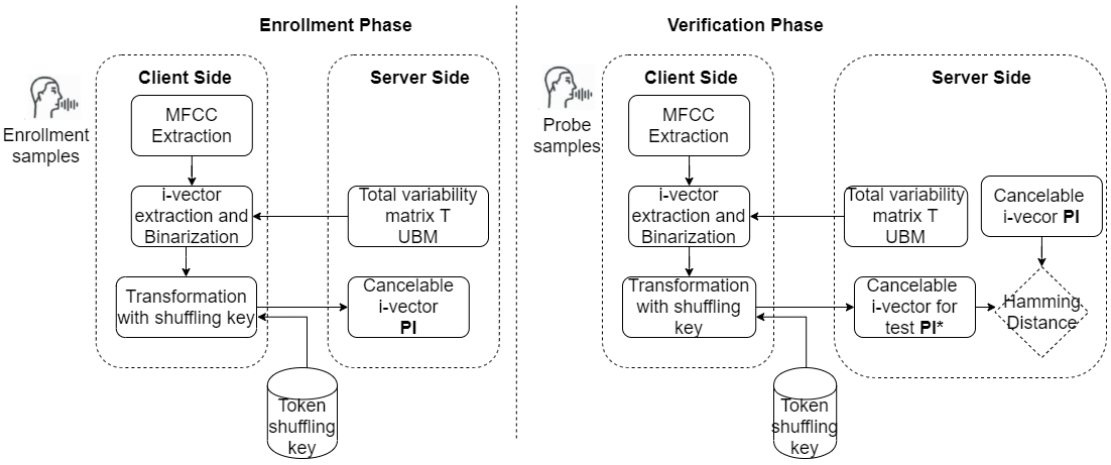


Figure 5.1: Architecture of the privacy-preserving speaker verification system based on cancelable i-vectors.

the client-side transforms the binary i-vector using the user's shuffling key received from the token and sends it to the server. As an output, the server receives the protected i-vector (cancelable i-vector) called the pseudonymous identifier PI.

During the verification phase, as an input, the user provides the probe voice samples to extract the MFCC features and the server has the total variability matrix T with the UBM and the pseudonymous identifier PI. The server sends the T and the UBM to the client-side to extract the test binary i-vector. Then the token sends the shuffling key to the client-side that transforms the binary i-vector with the shuffling key and generates the probe cancelable i-vector $PI*$. Finally, the probe cancelable i-vector is transferred to the server that measures the Hamming distance between the stored PI and the $PI*$ to decide the outcome of the verification.

Based on this protocol, the server never has access to the voice recorded by the user, and it does not possess a speaker's model in plaintext that could be misused. The server stores only the cancelable i-vector generated during enrollment and the total variability matrix T, that it does not reveal personal sensitive information.

5.3 Experimental Evaluation and Results

In this section, we evaluate the cancelable i-vector system according to the requirements described in the standard for the biometric information protection. Also, we demonstrate the feasibility of the proposed cancelable scheme to protect DNN speaker embeddings such as x-vectors. Furthermore, a security analysis of the cancelable i-vector system based on the evaluation methodology described in [Rosenberger, 2018] is reported.

5.3.1 Databases

For the evaluation of speaker verification system based on cancelable i-vectors, text-dependent and text-independent databases are used to study the feasibility of applying the cancelable scheme for both scenarios.

For text-dependent scenario, we report the biometric performance using the RSR2015 [Larcher et al., 2014] text-dependent database described in Chapter 4, section 4.3.1. For text-dependent scenario, we will show that due to the revocability property of the cancelable system, in case the passphrase of the target user is compromised, instead of selecting a new one, we can generate a new speaker biometric reference from the same compromised passphrase.

For text-independent scenario, we report the biometric performance using NIST 2016 Speaker Recognition Evaluation data [Sadjadi et al., 2017]. The dataset comprises utterances in two languages, Tagalog and Cantonese. Enrollment files have nominal durations of 60 s of speech whereas the duration of test files ranges from 10 to 60 s. The test set is composed of 37062 targets and 194 966 non targets trials for the pooled (female + male) condition.

In addition, we demonstrate that the proposed cancelable scheme could operate on DNN x-vecotors embeddings [Snyder et al., 2018] using VoxCeleb [Nagrani et al., 2017] text-independent database. VoxCeleb includes two datasets. VoxCeleb1 contains over 100,000 utterances for 1251 celebrities, while VoxCeleb2 contains over 1 million utterances for over 6112 celebrities extracted from videos uploaded to YouTube. The datasets are fairly gender-balanced, (VoxCeleb1 55% male, VoxCeleb2 61% male) and the speakers span a wide range of different eth-

nicities, accents, professions, and ages.

5.3.2 Experimental setups

For the evaluation of cancelable i-vector in text-dependent scenario on RSR2015 database, the MSR Identity toolbox [Sadjadi et al., 2013] was used. For speech features, 20-dimensional MFCCs are extracted with their first and second derivative and the log energy, leading to a 63-dimensional feature vector. Then, gender-independent UBM containing 1024 Gaussian is trained by using all male and female data of background partition of RSR2015 database. The UBM training data are reused for the training of the total variability matrix T of rank 400 by using 10 iterations of the expectation-maximization algorithm. After training the UBM and T, 400-dimension i-vectors are extracted. The i-vectors are passed through a linear discriminant analysis reducing their dimension from 400 to 200. The LDA is trained using the RSR2015 training data. Sentences having the same pass-phrase of a particular speaker are treated as belonging to the individual speaker class. This gives a total of $(50male + 47female) * 30 = 2910$ speaker-passphrase classes. To generate the cancelable i-vectors, the extracted i-vectors are binarized and transformed with the shuffling scheme. The cancelable speaker verification system is evaluated using the i-vectors without LDA and i-vectors passed through the LDA. Therefore, shuffling keys of lengths 400 and 200 are used to transform the binary i-vectors of dimensions 400 and 200 respectively.

For the evaluation of cancelable i-vector in text-independent scenario on SRE16 database, we use the the recipe available on Kaldi[1]. For speech features, 24-dimensional MFCCs are extracted with a frame length of 25 ms every 10 ms. These feature vectors are mean-normalized over a sliding window of up to 3 seconds. Then, energy-based voice activity detection (VAD) is applied to estimate frame-by-frame speech activity, and filter out non-speech frames. A UBM with 2048 mixture components is trained with the development SRE16 data. Then, i-vector extractor is trained with NIST SRE 2004-2010 and Switchboard databases. The i-vectors of dimension 600 are extracted for the test set of SRE16 and processed with mean subtraction, length normalization and LDA reducing their dimension from

[1],https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v1

600 to 200. The i-vectors are then binarized and transformed with the shuffling scheme to generate the cancelable templates.

For speaker verification based on x-vector system, we use the recipe available on Kaldi[2], where 512-dimensional x-vector speaker embeddings are extracted using a Time Delay Neural Network [Snyder et al., 2018] trained on Voxceleb2 [Chung et al., 2018] (dev and test portions) and the training portion (Dev) of VoxCeleb1 [Nagrani et al., 2017]. The x-vectors are binarized and transformed with the shuffling scheme to generate the cancelable x-vectors. The test set of VoxCeleb1 is used for the evaluation.

5.3.3 Biometric performance evaluation

The goal behind the biometric performance evaluation of speaker verification system based on cancelable i-vectors is to validate that the proposed protection scheme allows the protection of i-vectors without degradation in terms of biometric performance compared to the baseline system (without protection). Our goal is not to develop a cancelable system that outperforms the performance of the state of the art.

5.3.3.1 Biometric performance evaluation of the cancelable i-vector system on RSR2015 text-dependent database

Biometric performance of speaker verification systems based on the baseline, binary, and cancelable i-vectors in terms of EER are reported in Tables 5.1 and 5.2 on the female and male evaluation subset of part1 RSR2015. The evaluation was conducted under the legitimate scenario for target-correct, impostor-correct, and impostor-wrong trials of the RSR2015 database.

For the baseline system evaluation, the performance obtained was consistent with the results reported in [Larcher et al., 2014]. The speaker verification systems perform better with the impostor-wrong trials than impostor correct trials in terms of EER%. In fact, the impostor-correct trials are more challenging, as the non-target user pronounces the expected passphrase that is used to enroll the target

[2]https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2

Table 5.1: Biometric performance of the speaker verification systems based on the baseline, binary, and cancelable i-vectors. The performance is reported on the RSR2015 female evaluation subset for the impostor correct and impostor wrong trials in terms of EER (%).

System		**Baseline i-vector**	**Binary i-vector**	**Cancelable i-vector**
i-vector without LDA	tar-c/imp-c	2.55	7.27	1.12
	tar-c/imp-w	0.37	2.48	1.01
i-vector with LDA	tar-c/imp-c	3.39	9.43	0.08
	tar-c/imp-w	0.2	1.45	0.07

Table 5.2: Biometric performance of the speaker verification systems based on the baseline, binary, and cancelable i-vectors. The performance is reported on the RSR2015 male evaluation subset for the impostor correct and impostor wrong trials in terms of EER (%).

System		**Baseline i-vector**	**Binary i-vector**	**Cancelable i-vector**
i-vector without LDA	tar-c/imp-c	6.22	10.35	2.35
	tar-c/imp-w	2	4.84	2.5
i-vector with LDA	tar-c/imp-c	5	10.15	0.22
	tar-c/imp-w	0.4	0.92	1.04

speaker. As example, on the female subset, for target-correct / impostor-wrong trials, the EER=0.37% that increases to 2.55 % for target-correct / impostor-correct trials. Moreover, for the baseline i-vectors, we observe that the LDA performs better when applied to the i-vectors extracted from male speakers than from females. This could be explained by the fact that the number of speech utterances for males used to train the LDA was higher than the female ones. In addition compared to the performance reported in Chapter 4, Table 4.1, we observe that the system based on GMM performs better than the system based on i-vector since the GMM is more dedicated to text-dependent scenario.

For the speaker verification system based on cancelable i-vectors, the biometric performance outperforms that of the baseline (unprotected) i-vectors system. For example, for target-correct/impostor-correct trials on the female subset, the EER goes from 2.55% using the baseline i-vectors without LDA to 1.12% using the cancelable i-vectors. The proposed cancelable scheme improves the biometric

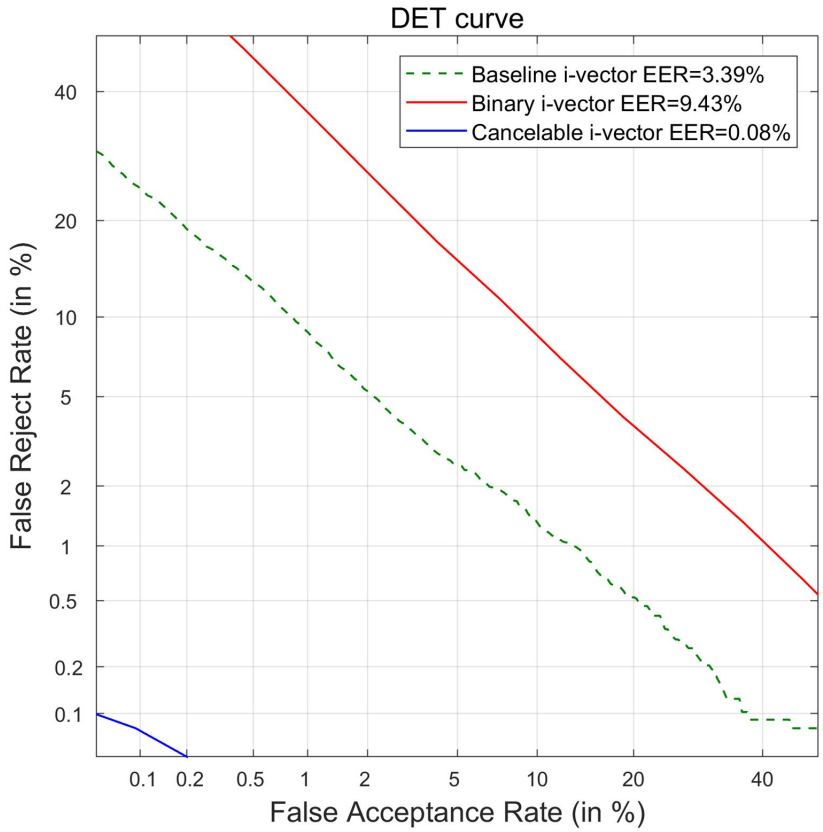


Figure 5.2: DET Curves for speaker verification systems based on the baseline i-vectors (with LDA) and the cancelable i-vectors using target-correct/impostor-correct trials on the female evaluation subset of RSR2015.

performance. As shown in Figure 5.2, the EER for speaker verification system based on the i-vectors with LDA is 3.39%, which improved to 0.08% using the cancelable i-vectors. For target-correct/impostor-wrong trials, the performance of cancelable i-vectors is close to that of the baseline i-vectors.

In addition, from the reported results we observed that the best performance of the cancelable system was obtained using the cancelable i-vectors extracted from the i-vectors passed through the LDA. This could be explained through Figure 5.3, where it is shown that the overlap between the target and non-target scores distributions of cancelable i-vectors with LDA is smaller than the one obtained without LDA. In fact, before applying the shuffling, the mean of the distribution of the non-target scores using the binary i-vectors is 0.35 with i-vectors passed through

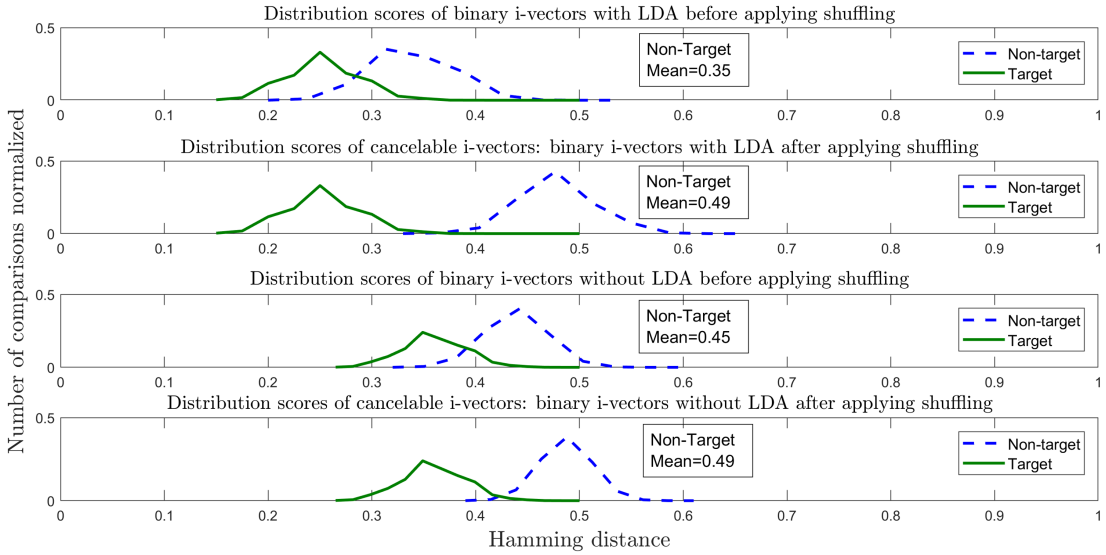


Figure 5.3: Distribution scores of target correct/impostor-correct trials on the female evaluation subset of RSR2015.

the LDA and 0.45 using i-vectors without LDA. When applying the shuffling to transform the binary i-vectors with LDA, the mean is shifted by 0.14 and moves from 0.35 to 0.49, resulting in a separation between the target and no-target distributions. However, for binary i-vectors without LDA, the mean is only shifted by 0.04 from 0.45 to 0.49 which does not allow a good separation between target and no-target scores.

5.3.3.2 Biometric performance evaluation of the cancelable i-vector system on SRE16 text-independent database

Table 5.3 reports the biometric performance of the speaker verification systems based on the baseline, binary, and cancelable i-vectors in terms of EER on the evaluation SRE16 database. As a baseline system, we use the recipe available on Kaldi 1, where 600-dimensional i-vector are extracted and processed with mean subtraction, length normalization, and LDA reducing their dimension from 600 to 200. For the cancelable system, the 200-dimensional i-vector is binarized using the median and transformed with a shuffling key of dimension 200.

As shown in Figure 5.4 and Table 5.3, the protection of the i-vector with the

shuffling scheme does not degrade the biometric performance compared to the baseline (unprotected) system. Speaker verification system based on cancelable i-vectors maintains the performance with EER=12.57% compared to the baseline i-vectors with cosine scoring as back-end (EER=16.74%).

In Table 5.3, we report the biometric performance of the speaker verification

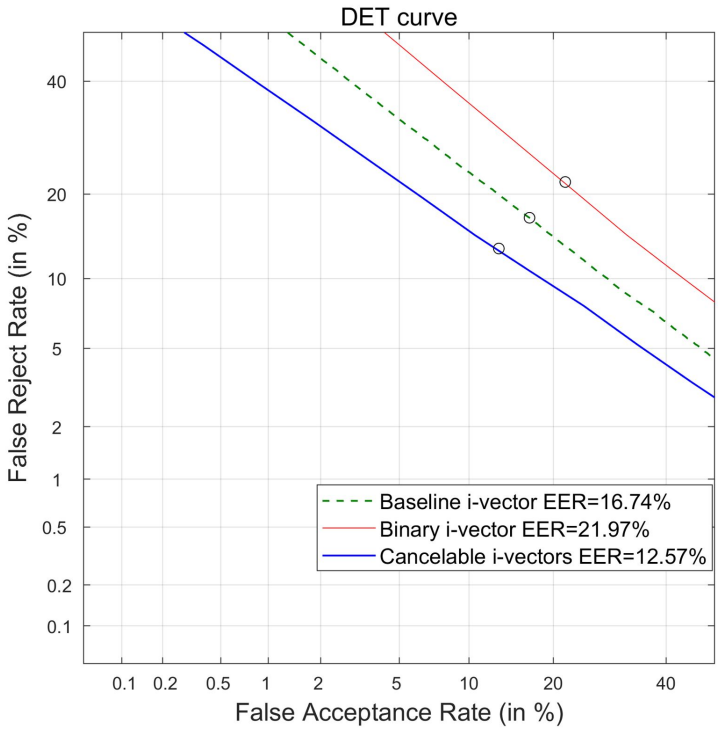


Figure 5.4: DET Curves for speaker verification systems based on the baseline (non-protected) i-vectors, binary, and the cancelable i-vectors on the evaluation set of SRE16 database.

Table 5.3: Biometric performance of the speaker verification systems based on the baseline, binary, and cancelable i-vectors on the text-independent SRE16 evaluation database.

Systems	**Back-end scoring**	**EER%**
Baseline i-vectors	Cosine	16.74
Baseline i-vectors + Shuffling	Cosine	5.88
Binary i-vectors	Hamming	21.97
Cancelable i-vectors	Hamming	12.57

system based on i-vectors transformed directly with the shuffling scheme without passing by the binarization step (baseline+shuffling). As reported, the EER obtained equal to 5.88% is better than the performance reported with shuffled binary i-vectors (cancelable i-vectors). However, without binarization, in case the shuffling key is compromised, the original i-vector will be recovered since the shuffling scheme is reversible. The goal behind the binarization is to hide the original i-vector.

Based on the above evaluation, we can conclude that the protection of i-vectors passed through the LDA or not with the cancelable scheme (binarization + shuffling) maintains the biometric performance compared to the unprotected (baseline) i-vectors in terms of verification accuracy.

5.3.3.3 Biometric performance evaluation of the cancelable x-vector system on the VoxCeleb text-independent database

The proposed cancelable scheme shows its effectiveness on the speaker verification based on i-vector. In order to demonstrate its feasibility on the state-of-the-art speaker verification systems, we used this cancelable scheme to protect x-vectors speaker embeddings. As a baseline x-vector system, we adopt the recipe available on Kaldi1 where 512-dimensional x-vector speaker embeddings are extracted using a Time Delay Neural Network [Snyder et al., 2018] trained on VoxCeleb1,2 [Nagrani et al., 2017]. For the back-end scoring, we use simple cosine scoring without normalization and dimensionality reduction. For the cancelable system, the 512-dimensional speaker's x-vector is binarized using the median as described in subsection 5.2.1 and then it is transformed with the shuffling scheme. The comparison is performed with Hamming distance.

Results in Table 5.4 validate that the shuffling scheme allows the protection of the x-vectors without a degradation in terms of EER. As shown, cancelable x-vectors perform better with EER=0.05% than the baseline x-vectors with cosine scoring as back-end (EER=8.18%) and even better than x-vectors results reported in Kaldi recipe with PLDA as back-end scoring (EER=3.12%) [Snyder et al., 2018]. However, the proposed cancelable scheme is not dedicated to protect speaker verification system based on log-likelihood scores because it does not take into

Table 5.4: Biometric performance of the cancelable and the baseline x-vector systems on the test set of VoxCeleb text-independent database in terms of EER (%).

System	Back-end scoring for the baseline	Baseline x-vector	Binary x-vector	Cancelable x-vector
x-vector	Cosine	8.18	9.68	0.05
	PLDA [Snyder et al., 2018]	3.12		

consideration the protection of PLDA model parameters. Otherwise, during the latest NIST SRE'19 speaker recognition evaluation, the x-vectors extracted from residual networks using cosine distance scoring performed the best on the VAST database avoiding the need for PLDA [Villalba et al., 2020]. We believe that the proposed cancelable scheme could be applied to protect such state-of-the-art systems.

5.3.4 Revocability analysis of the cancelable i-vector system

As described in the revocability analysis of cancelable GMM system in Chapter 4, subsection 4.3.5, revocability is evaluated by calculating the pseudo-impostor scores. For this, we shuffled one speaker's binary i-vector with 480000 randomly generated shuffling keys. The first shuffled binary i-vector is compared with the remaining shuffled templates to compute the pseudo-impostor scores. This process is repeated with 30 different users.

From Figure 5.5, we can notice that the distribution of the pseudo-impostor scores overlaps with the distribution of the non-target scores. This indicates that the newly generated cancelable i-vectors are indistinguishable, although they are generated from the same binary i-vector. Since the newly generated cancelable templates are uncorrelated, this justifies that the system achieves the revocability requirements. Therefore, when the passphrase is compromised in a text-dependent speaker verification system, instead of selecting a new one, we can generate a new speaker reference from the compromised passphrase.
We estimate the maximum number of possible cancelable i-vectors that can be generated from the same binary i-vector using the Hamming bound [MacWilliams and

Sloane, 1977] as described in Chapter 4, section 4.3.5. Using the EER threshold $t = 0.4$ of the cancelable system, for i-vector and shuffling key of length 400-bits, we get almost 2^{12} possible cancelable i-vectors PI for each user as given in Eq. 5.4.

$$PI = \frac{Number\ of\ possible\ permutation}{Volume\ of\ Hamming\ spheres} = \frac{400!}{(200!)(200!)\sum_{k=0}^{(t\times 400)}\binom{400}{k}} \approx 2^{12} \tag{5.4}$$

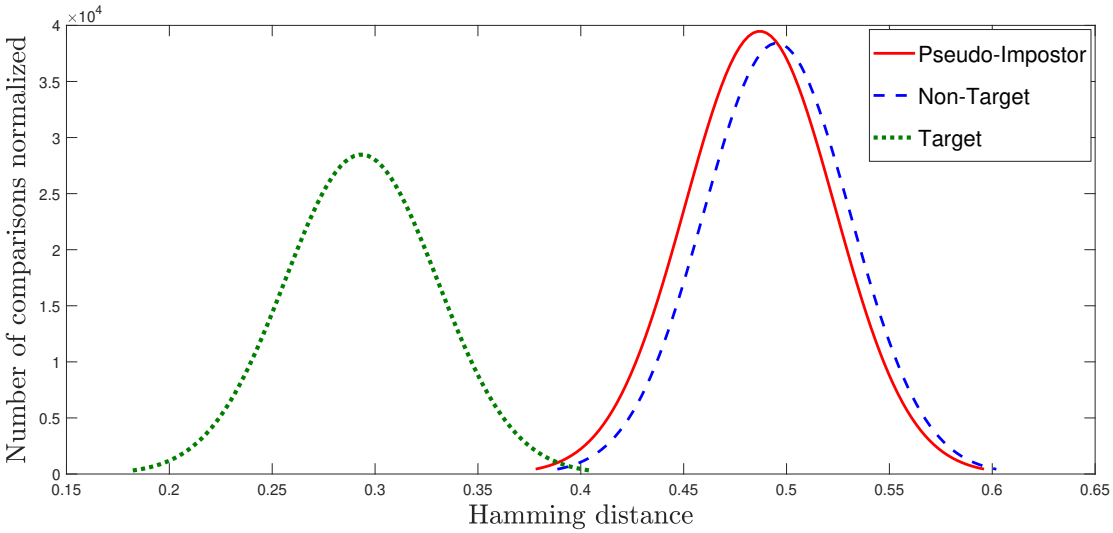


Figure 5.5: Revocability analysis of the cancelable i-vector system: Distribution of Target, Non-Target and Pseudo-impostor scores on the female evaluation subset (target-correct/imposor-wrong trials) of RSR2015.

5.3.5 Unlinkability analysis of the cancelable i-vector system

The unlinkability of cancelable i-vectors is evaluated based on the framework described in [Gomez-Barrero et al., 2017]. Therefore, two types of scores are computed. *Mated instances:* scores computed by comparing cancelable i-vectors extracted from different samples of the same subject using different shuffling keys. *Non-mated instances:* scores computed by comparing cancelable i-vectors gener-

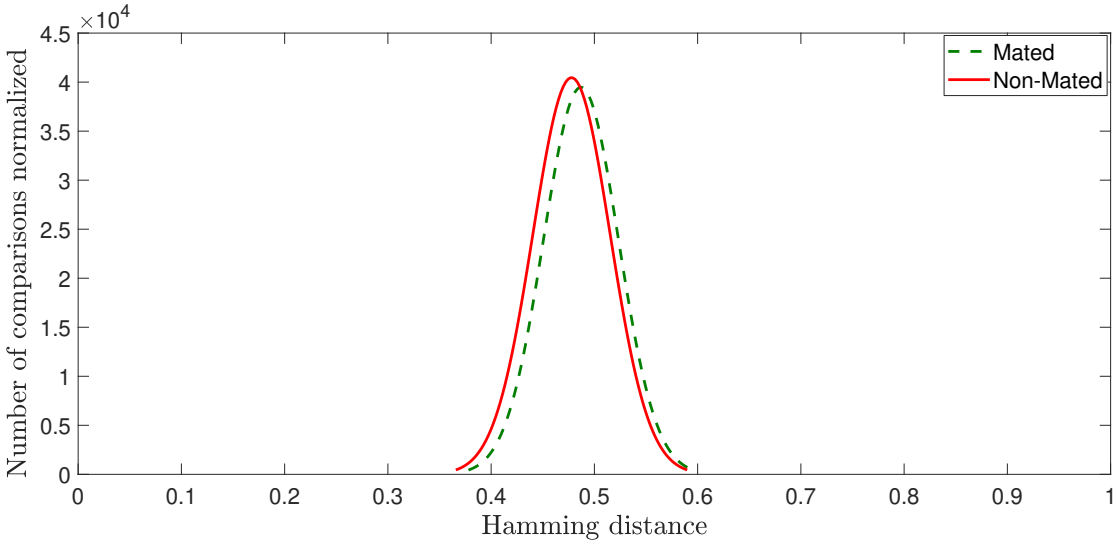


Figure 5.6: Unlinkability analysis of the cancelable i-vector system: distribution of Mated and Non-mated scores using the female subset of RSR2015 database.

ated from samples of different subjects using different shuffling keys.

As we explained in chapter 4, subsection 4.3.6, the global metric $D_{\leftrightarrow}^{sys}$ gives an estimation of the global linkability of the system. As observed in Figure. 5.6, the distribution of mated and non-mated scores are overlapped with global linkability $D_{\leftrightarrow}^{sys}$ equal to 0 rendering the system fully unlinkable.

5.3.6 Irreversibility analysis of the cancelable i-vector system

The irreversibility analysis of cancelable i-vector based on shuffling scheme is the same as that of the cancelable GMM proposed in Chapter 4. Using the shuffling key and the cancelable i-vector, the reconstruction of binary i-vector representation is possible. However, due to the binarization step, it is not possible to recover the original i-vector.

In case the attacker has only the cancelable template without the shuffling key, if the adversary wants to guess the correct values of the binary i-vector of length 400-bits, it is computationally not feasible. In fact, the guessing complexity is

huge and equal to 2^{395} the number of possible permutation, given by Eq. 5.5.

$$Number\ of\ possible\ permutation = \frac{400!}{(200!)^2} \approx 2^{395} \tag{5.5}$$

5.3.7 Security analysis of the cancelable i-vector system

The proposed cancelable i-vector is based on two factors, the biometric sample and the shuffling key. Herein, we report the robustness of this cancelable system in case these factors are compromised. For this, we follow the methodology proposed in [Rosenberger, 2018], proposing different attacks to evaluate the security of cancelable systems:

Zero effort attack A_1*:* Non-target user provides his/her biometric samples and a random shuffling key to impersonate the target user.

Brute force attack A_2*:* Non-target user tries to be verified by trying different random values of cancelable i-vectors.

Stolen token attack A_3*:* Non-target user has stolen the shuffling key of the target user and tries random binary vectors to generate the target's cancelable i-vector.

Stolen biometric data attack A_4*:* Non-target user has stolen the biometric samples of the target user and tries with random shuffling keys to generate the target cancelable i-vector.

Worst case attack A_5: Non-target user has stolen the target user's shuffling key and provides its own biometric samples to generate the cancelable i-vector.

Very worst case attack A_6: Non-target user has stolen the target shuffling key and has a wrong pass-phrase spoken by the target user. This attack is specified for text-dependent scenario.

For these attack scenarios, we compute the false acceptance rate for each attack scenario A_i, when the EER threshold of the cancelable i-vector system ε_{EER} in the legitimate scenario is taken as the decision threshold. A high value of A_i implies that the system is not robust to this attack scenario. Table 5.5 presents the values of A_i obtained when the cancelable i-vector system is attacked with scenarios A_1, A_2, A_3, and A_4. Also, in Figure 5.7, we present the evolution of the FAR curve for

Table 5.5: Security evaluation in terms of FAR reported at the EER threshold ε_{EER} of the cancelable i-vector system.

Attack scenario	**FAR** (ε_{EER})
Zero effort attack A_1	0
Brute force attack A_2	0
Stolen token attack A_3	0
Stolen biometric data attack A_4	0

each attack scenario related to the EER threshold of the cancelable i-vector in the legitimate scenario. As shown, the cancelable system is robust for all presented attack scenarios with FAR=0 at the EER threshold.

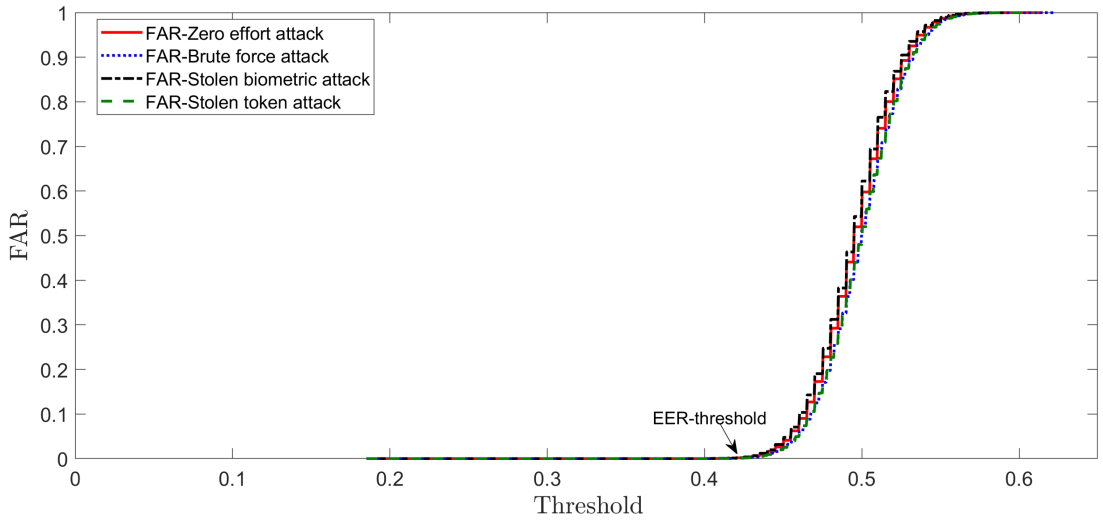


Figure 5.7: Evolution of the FAR curves for the cancelable i-vector system against the attack scenarios A_1, A_2, A_3, and A_4 using the female evaluation subset of RSR2015 database.

For the worst-case scenario A_5, we evaluate the robustness of the system in two scenarios:

Worst-case scenario 1: the attacker provides the target user's shuffling key and pronounces the wrong pass-phrase.

Worst-case scenario 2: the attacker provides the target user's shuffling key and pronounces the correct pass-phrase.

Table 5.6: FAR of the worst case scenario 1 in case the attacker pronounces the wrong pass-phrase. The FAR is reported according to the performance of cancelable i-vector system in terms of FAR and FFR in the legitimate scenario

Legitimate scenario	**FAR %**	1.01	0.7	0.4	0.2	0.1
	FRR %	1.01	1.4	1.9	2.88	3.9
FAR-worst case % Scenario 1		7.12	4.88	3.2	2.08	1.3

Table 5.7: FAR of the worst case scenario 2 in case the attacker pronounces the correct pass-phrase. The FAR is reported according to the performance of cancelable i-vector system in terms of FAR and FFR in the legitimate scenario

Legitimate scenario	**FAR %**	1.12	0.63	0.35	0.18	0.02
	FRR %	1.12	1.4	1.9	2.88	7.27
FAR-worst case % Scenario 2		27.55	22	17.21	13.82	7.27

Tables 5.6 and 5.7 report the FAR obtained for the two worst-case scenarios. Also, in Figure 5.8 we present the FAR curves for the legitimate and the worst-case scenarios. Results show that an acceptable FAR is obtained for the worst-case scenario 1. At the threshold corresponding to FAR=0.4% and FRR=1.9% in the legitimate scenario, the FAR=3.2% for the worst-case attack. However, in case the attacker pronounces the correct pass-phrase (worst-case scenario 2) a degradation in FAR is observed. For FAR=0.35% and FRR=1.9% in the legitimate scenario, the FAR=17.21% for the worst-case attack. In fact, in this scenario, the same shuffling key is used to transform the binary i-vector representations of the target and non-target users. Therefore, the biometric performance of cancelable i-vectors will be the same as obtained using the unprotected binary i-vectors. Regarding results reported in Table 5.1, for target-correct/impostor-correct trials, the performance at binary level degrades further comparing to impostor wrong trials which explain the degradation in the FAR for the worst-case scenario 2.

For the very worst case attack A_6, where the attacker provides a wrong-pass-phrase spoken by the target user and the target shuffling key, the FAR=10.4% at the EER threshold of legitimate scenario.

Based on this security analysis, we can conclude that the proposed cancelable

i-vector system, as well as the cancelable GMM, are robust to A_1, A_2, A_3, and A_4 attack scenarios. However, for the worst-case attack, a degradation in terms of false acceptance rate was observed. This degradation is related to the loss of biometric performance when transforming the speaker's i-vector or GMM into a binary vector. The security of the shuffling key for the proposed system is very important. We believe that in real use cases such security can be guaranteed with the novel technologies as the Embedded Secure Element [Tremlet, 2016] or the secure chip which provides a secure space to store and manage personal data.

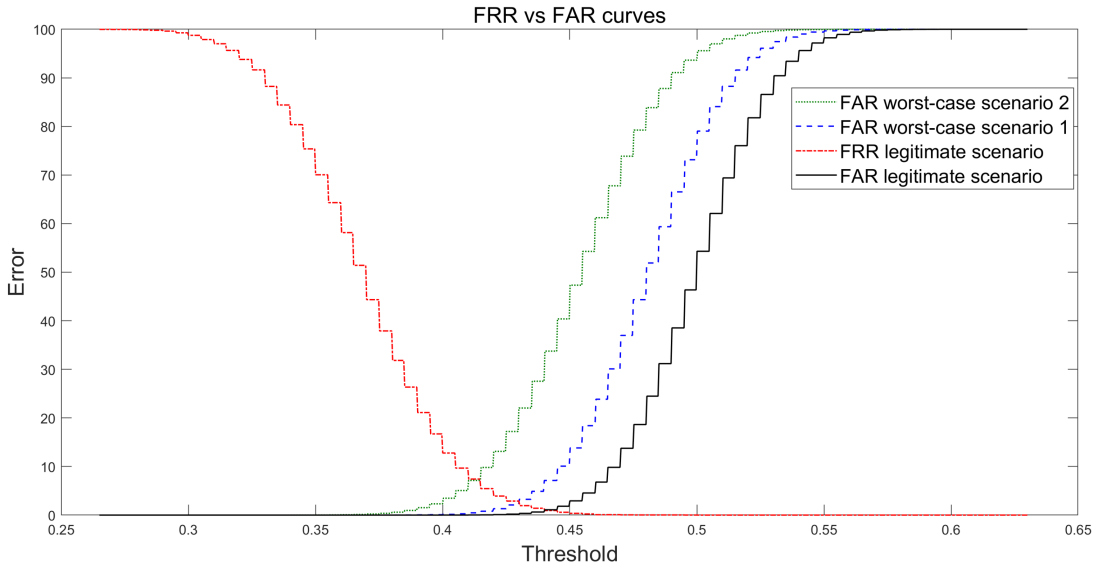


Figure 5.8: Evolution of the FAR curves for the cancelable i-vector system for the legitimate scenarios and the worst-case scenarios A_5 using the female evaluation subset of RSR2015.

5.4 Chapter Summary and Conclusions

In this chapter, we proposed a cancelable scheme for privacy-preserving speaker verification systems based on i-vectors. This is achieved by first binarizing the i-vector and then its transformation with the shuffling scheme. We also demonstrate that this cancelable scheme could operate to protect speaker verification systems based on deep neural network speaker embeddings such as x-vectors.

In order to make our research reproducible, the cancelable scheme was evaluated using public databases following a clear protocol. The RSR2015 text-dependent database was used to evaluate the system based on i-vectors and the VoxCeleb text-independent database for the system based on x-vectors. The main findings of this chapter can be summarized in the followings points:

- We propose a cancelable speaker verification system to mitigate privacy and security issues based on two steps; the i-vector binarization by thresholding the i-vector with its median value, and then the transformation of the binary i-vector with the shuffling scheme.
- The speaker verification system based on the cancelable i-vectors reaches better biometric performance than the baseline i-vector system contrary to existing privacy protection methods.
- The cancelable i-vectors system achieves the biometric information protection requirements [ISO/IEC JTC1 SC27 Security Techniques, 2011], and shows a good level of security against different attack scenarios.
- We also demonstrate that this cancelable scheme could operate on the state-of-the-art speaker verification systems based on Deep Neural Network (DNN) speaker embeddings.

However, the main weakness of the cancelable i-vector system is its low resistance to the worst-case attack (stolen key scenario), which is related to the degradation of biometric performance caused by the transformation of the speaker's i-vector into a binary representation. In chapter 6, a novel approach for the binarization of speaker representation while maintaining the biometric performance will be presented.

Chapter 6

Privacy-Preserving Speaker Verification System Based on Cancelable x-Vectors

Contents

In this chapter, we have improved the privacy-preserving scheme presented in Chapter 5 in order to resolve the shortcomings regarding the degradation of biometric performance during the binarization and its impact on the robustness of the system against stolen key attacks. We present a new privacy-preserving scheme where we propose a binarization approach of speaker biometric reference that maintains the biometric performance. Also, we propose to combine the shuffling scheme with secure sketch error correction. We will show that applying secure sketch error correction to cancelable biometrics improves the biometric performance and the robustness against stolen key attacks. The proposed protection scheme could be applied for speaker verification systems based on i-vector or x-vector. In this chapter, the recent speaker verification system based on x-vector embeddings is taken as the baseline (unprotected) system and we will use the proposed protection scheme to develop a privacy-preserving speaker verification system based on cancelable x-vectors.

The proposed system includes three main stages: (i) x-vector extraction and binarization, (ii) extraction of cancelable x-vector by transforming the binary x-vector with the shuffling scheme, and (iii) applying the secure sketch to the cancelable x-vector by passing it through an error-correcting code. The proposed system is evaluated according to the requirements of biometric information protection [ISO/IEC JTC1 SC27 Security Techniques, 2011] in terms of biometric performance, revocability, irreversibility, and unlinkability. Also, the robustness against different attack scenarios was analyzed.

The chapter is structured as follows. Section 6.1 gives a general presentation of the proposed system including a description of enrollment and verification phases. Section 6.2 presents the baseline (unprotected) x-vectors speaker verification systems and its biometric evaluation. Section 6.3 describes the proposed binarization approach of the x-vectors embeddings and presents the biometric performance evaluation of the speaker verification system based on binary x-vectors. Section 6.4 presents the description and the evaluation of the cancelable x-vector which is the output of the transformation of the binary x-vector with the shuffling scheme. In section 6.5, we describe the application of secure sketch error correction to cancelable x-vectors and we present the evaluation of biometric performance, privacy requirements, and security analysis. The chapter summary and conclusion

are presented in section 6.7.

6.1 Privacy-Preserving Speaker Verification System Based on x-Vectors

The main weakness of the privacy-preserving system presented in chapter 5 is the low resistance to stolen key attacks, which is related to the degradation of biometric performance caused by the binarization of the speaker model. Therefore, we propose for the new privacy-preserving scheme:

1. A Binarization approach of the speaker model that maintains the biometric performance of the baseline system.

2. Applying secure sketch error correction code (ECC) to the binary x-vectors transformed with the shuffling scheme. The idea is to pass the cancelable x-vector through an error-correcting code to manage the biometric variability which allows to improve the false acceptance rate in the stolen key scenario.

In this section, we present an overview of the proposed privacy-preserving speaker verification system based on x-vectors, including descriptions of the enrollment and verification phases.

6.1.1 Enrollment phase

The enrollment phase includes two steps, i) extraction of the secure sketch error correction corresponding to each speaker and ii) extraction of the speaker's cancelable x-vector for the enrollment.

Step 1 of enrollment: Secure sketch extraction
The sketch scheme defined by Dodis *et al.* [Dodis et al., 2004] consists of two algorithms: a sketch generation algorithm Gen, and a reconstruction algorithm Rec. Given some data X the output $P_X = Gen(X)$ is called a sketch of X. Given a sketch P_X and another Y that is sufficiently similar to X according to some measure, $Rec(P_X, Y)$ would reconstruct the original X. For biometric system,

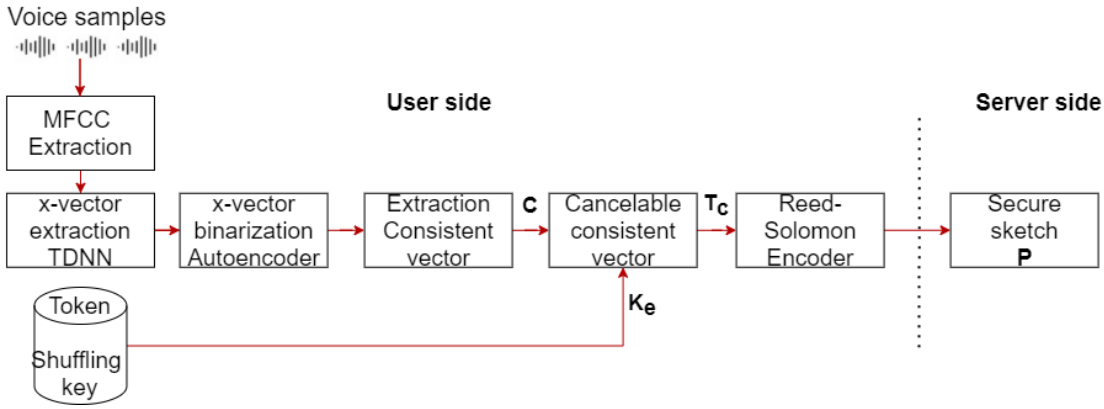


Figure 6.1: Pipeline for Step 1 of enrollment: Extraction of secure sketch.

secure sketch consists on the reconstruction of a biometric input X with the help of the secure sketch of X, P_X and a noisy biometric input Y closer to X.

During step 1 of enrollment of the proposed system, the *Gen* algorithm will be used to generate for each user a secure sketch. For that, as shown in Figure 6.1, the user provides voice samples from which the system first extracts the MFCC features and then the corresponding x-vectors embeddings for each sample using a trained Time Delay Neural Network (TDNN) model. The TDNN used to extract the x-vectors is described in section 6.2. Next, the x-vectors are transformed into binary embeddings B using a binarization method based on an autoencoder model. This binarization method is described in section 6.3.

Then, a consistent vector C representing the user is extracted by considering the significant bits in his/her binary embeddings. Consistent bits are those bits in the binary embeddings which are less likely to change. Consistent vector bits are derived after aligning and summing up the binary embeddings to examine the occurrence of bits. Then, a bit is set to one in C if the probability of occurrences is greater or equal to a specific threshold p_{th} across the binary x-vectors, if not it takes 0 as defined in Eq 6.1 and Eq 6.2:

$$C(i) = \begin{cases} 1 & for \quad p(i) \geq p_{th} \\ 0 & elsewhere \end{cases} \tag{6.1}$$

$$p(i) = \frac{\sum_{r=1}^{R} B_r(i)}{R} \tag{6.2}$$

where R is total number of samples or binary embeddings B and $p(i)$ is the probability of i_{th} bit in the binary embedding. In our work, we have chosen empirically the value of R=3 and the threshold p_{th}= 0.66.

The consistent binary x-vector is then transformed with the user specific shuffling key to generate his/her cancelable consistent template T_c and passed through the *Gen* algorithm to generate the secure sketch. The secure sketch is generated from the cancelable T_c. Therefore, in case it is compromised, a new one can be generated by transforming T_c with a new shuffling key.

The Reed-Solomon (RS) error correction code [MacWilliams and Sloane, 1977] has been implemented to extract the secure sketch. The goal was to combine the security of both shuffling scheme and EEC and take advantage of shuffling transformation and error correction to improve the biometric performance. For this, the cancelable consistent template T_c is passed through the RS code to generate a user specific-secure sketch which corresponds to the parity symbols P extracted from the RS encoding of T_c. At the end of this first step of enrollment, we store only the secure sketch P on the access control system and we delete the rest. A detailed description about the implementation of the Reed-Solomon code will be provided in section 6.5.

The goal behind this first step is to generate for each user a secure sketch which is the RS parity symbols from RS encoding of his/her cancelable consistent binary x-vector T_c. The process is summarized as follows:

i) The user provides three voice samples.

ii) Extraction of x-vectors corresponding to the three voices samples.

iii) Binarization of the x-vectors.

iv) Extraction of the consistent x-vector C from the three binary x-vectors.

v) Generation of the cancelable consistent x-vector T_c by transforming C with the user specific shuffling key.

vi) Generation of the secure sketch P by passing the cancelable T_c through the RS encoder.

vii) Store the secure sketch P and delete the rest.

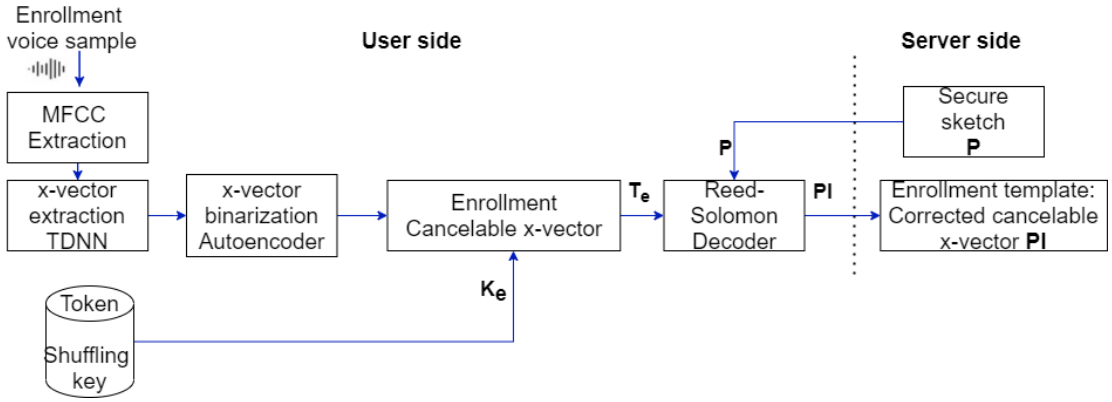


Figure 6.2: Pipeline for step 2 of enrollment: Extraction of the enrollment cancelable x-vectors.

Step 2 of enrollment: Extraction of enrollment cancelable x-vectors
During this step, the reconstruction algorithm Rec is used to extract the enrollment cancelable x-vector. As shown in Figure 6.2, the user provides its enrollment voice sample to the x-vector embedding extractor and binarization modules. The output is a binary x-vector representation. Then, the binary x-vector is transformed using the user-specific shuffling key K_e to generate the enrollment cancelable x-vector T_e. The binarization and the shuffling scheme allow to hide and protect the original x-vector and help in achieving revocability because if the cancelable x-vector or the shuffling key are compromised, a new cancelable template can be generated by the transformation of the binary representation with a new shuffling key.

Next, the cancelable x-vector T_e is passed through the Reed Solomon decoder. The RS decoder assumes that T_e is a noisy version of the cancelable consistent binary x-vector T_c, and takes T_e and the user secure sketch P received from the access system and performs the decoding. The RS decoding of T_e using P extracted during step 1 of enrollment allows to generate an enrollment cancelable x-vector close to the cancelable consistent template T_c representing this user. The result of decoding represents the corrected cancelable x-vector which represents the user enrollment template PI. After the enrollment phase, the access system has only the enrollment template PI and the secure sketch P. The process of enrollment step 2 is summarized as follows:

i) The user provides the enrollment voice sample.

ii) Extraction of the enrollment x-vector embedding.

iii) Binarization of the enrollment x-vector.

iv) Generation of the cancelable x-vector T_e by transforming the binary x-vector with the shuffling scheme.

v) Generation of the corrected cancelable x-vector PI by applying the RS decoding (Rec algorithm) to T_e using the secure sketch P of the claimed user.

vi) Store the enrollment template PI in the server-side.

6.1.2 Verification phase

During the verification, the user presents the probe voice sample and the shuffling key. The key could be the same as the enrollment key in the case of genuine access or it could be a random key in the case of impostor access. The probe sample is passed through the x-vector embedding extractor and binarization modules to obtain the probe binary x-vector. Using the shuffling key provided by the user, the probe cancelable x-vector T_p is generated. Then, the RS decoder assumes that T_p is an error-prone version of T_c, it combines the secure sketch P of the claimed identity with T_p, and performs the decoding process to reconstruct T'_p which represents the user's probe template PI^*.

The probe cancelable x-vector T_p could correspond to a genuine or impostor user. The error correction capability should be chosen so that the EEC can reduce the intra-variability while preserving the inter-variability. Thus in case T_p corresponds to a genuine user, the RS decoder will be able to reduce the variability of T_p and reconstruct T'_p closest to the enrollment cancelable template T_c. However, in case T_p corresponds to the impostor user, the RS decoder will not be able to reconstruct T'_p closest to the genuine enrollment template T_c. In fact, due to the application of the shuffling scheme that separates genuine and impostor score distributions, the distance between T_p and T_c exceeds the error correction capacity. The process of verification is summarized as follows:

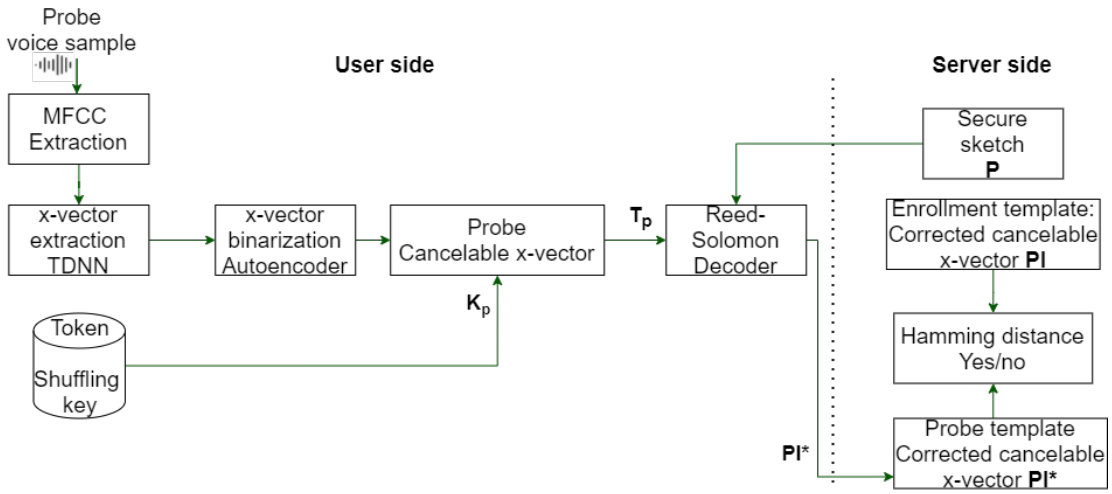


Figure 6.3: Pipeline for the verification phase of the proposed privacy-preserving speaker verification system based on cancelable x-vectors.

i) The user provides the probe voice sample.

ii) Extraction of the probe x-vector embedding.

iii) Binarization of the probe x-vector.

iv) Generation of the probe cancelable x-vector T_p by transforming the probe binary x-vector with the shuffling scheme.

v) Generation of the probe corrected cancelable x-vector PI^* by passing T_p combined with the secure sketch P of the claimed identity through the RS decoding.

vi) Compute the Hamming distance between the probe template PI^* and the enrollment template PI and compare it to the verification threshold.

In the following sections, detailed descriptions and evaluations of each module of the proposed privacy-preserving speaker verification system based on cancelable x-vectors are presented.

6.2 Baseline Speaker Verification System Based on x-Vector Embeddings

6.2.1 Description of the baseline speaker verification system based on x-vectors

Recently, researchers proposed end-to-end speaker recognition systems based on x-vectors embeddings [Snyder et al., 2016], [Snyder et al., 2017], and [Snyder et al., 2018]. In the end-to-end x-vector approach, deep neural networks are fed with a variable-length utterance and map it to speaker embedding. Figure 6.4 presents the basic structure of the DNN in the x-vector based system. The network is composed of three parts. First, an encoder network extracts frame-level representations from acoustic features such as the Mel-Frequency Cepstral Coefficients (MFCC). Then, a statistics pooling layer aggregates the frame-level representations into a single vector per utterance. Next, at segment-level a feed-forward classification network processes this single vector to calculate speaker class posteriors with softmax output layer [Goodfellow et al., 2016]. The x-vector embedding is extracted from the affine transform after the pooling layer.

Different x-vector systems are proposed in the literature characterized by different encoder architectures, pooling methods, and training objectives. For our work, we use the DNN embedding illustrated in Table 6.1 based on Time Delay Neural Network described in [Snyder et al., 2018].

The network consists of five layers that operate on speech frames, a statistics pooling layer that aggregates over the frame-level representations, two layers that operate at the segment level, and finally a softmax output layer. The first 5 layers of the network work at the frame level, with a time-delay architecture [Peddinti et al., 2015]. Each feature frame from a given utterance is captured by a sequence of time-delay layers. Suppose an input utterance has T frames, the time delay layers operate on speech frames with a small temporal context centered at the current frame t. For example, the input to layer frame3 is the spliced output of frame2, at frames t - 3, t and t + 3. The frame-level representation at each layer aggregates information from the context of previous layer, so that frame3 sees a total context of 15 frames. Then, the statistics pooling layer aggregates

all T frame-level outputs from layer frame5 and computes its mean and standard deviation that are aggregated and propagated through the segment-level layers and the softmax output layer.

The DNN is trained to classify the N speakers in the training data. After training, segment 7 and the softmax layer are removed and the x-vector embedding is extracted from the affine component of layer segment6. The x-vector is considered as the speaker biometric reference that will be used for the verification task.

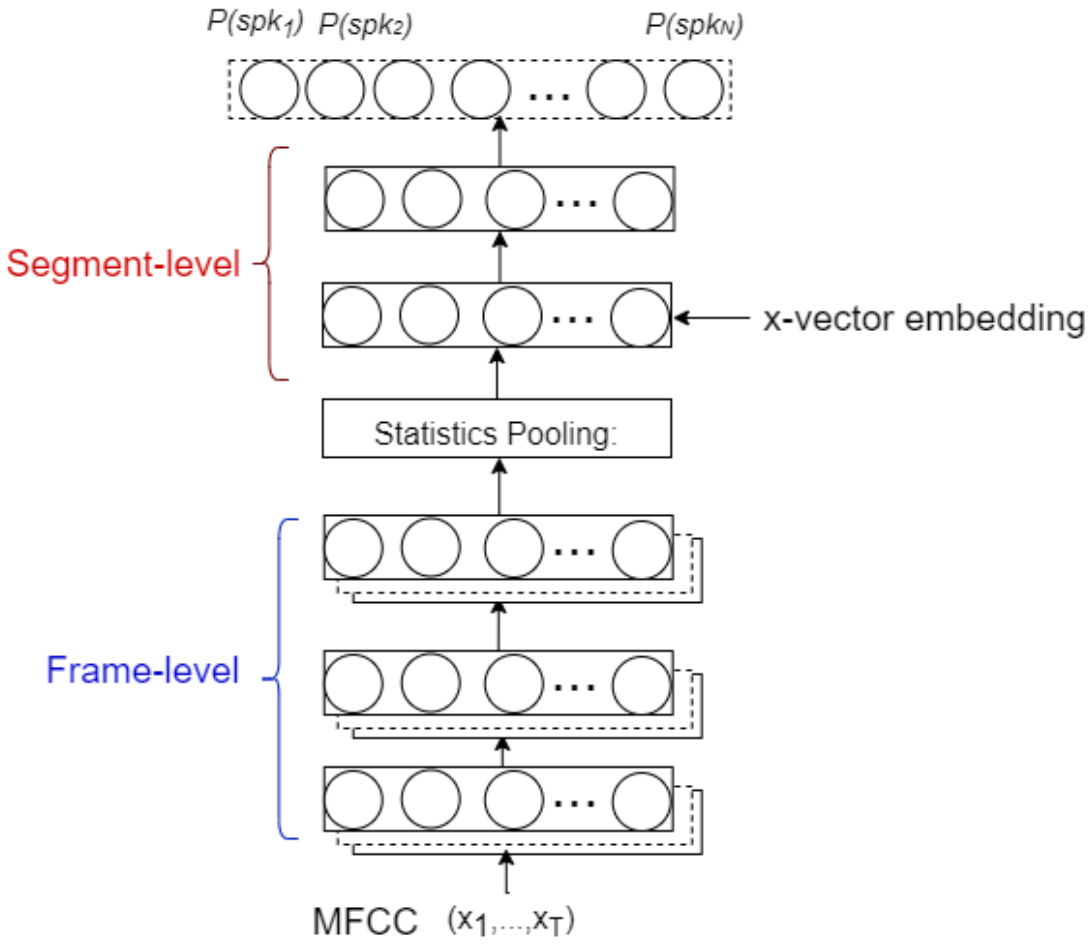


Figure 6.4: Structure of the DNN in the x-vector-based system. Frame-level operates on speech frames to extract frame-level representation. Statistics pooling layer aggregates all the frame-level outputs into a single vector and propagates it through the segment-level layers and the classification output layer [Snyder et al., 2017].

Table 6.1: Architecture of the DNN used for the extraction of x-vectors. The x-vectors are extracted at segment-level 6. T represents the number of frames in the input utterances and K is the number of parameters per frame. The N in the softmax layer corresponds to the number of training speakers.

Layer	Layer context	Total context	Input x output
frame-level1	[t - 2, t + 2]	5	Kx512
frame-level2	{t - 2, t ,t + 2}	9	1536x512
frame-level3	{t - 3, t ,t + 3}	15	1536x512
frame-level4	{t}	15	512x512
frame-level5	{t}	15	512x1500
stats pooling	[0, T)	T	1500Tx3000
segment-level6	{0}	T	3000x512
segment-level7	{0}	T	512x512
softmax	{0}	T	512xN

6.2.2 Experimental evaluation and results of the baseline x-vector speaker verification system

6.2.2.1 Experimental settings

For the baseline x-vector system, we use the recipe available on Kaldi speech recognition toolkit[1] [Povey et al., 2011]. For speech features, 24-dimensional MFCCs are extracted with a frame length of 25 ms every 10 ms. These feature vectors are mean-normalized over a sliding window of up to 3 seconds. Then, energy-based voice activity detection (VAD) is used to estimate frame-by-frame speech activity, and filter out nonspeech frames.

The TDNN is trained using 1 276 888 utterances from 7 323 speakers of the text-independent databases Voxceleb2 [Chung et al., 2018] (dev and test portions) and the training portion (Dev) of VoxCeleb1 [Nagrani et al., 2017] collected from celebrities videos uploaded to YouTube. Besides, as suggested in [Snyder et al., 2018], data augmentation was performed to increase the amount and diversity of the available training data. The augmentation strategy was used to add four corrupted copies of the original recordings to the training list. The recordings are corrupted by employing additive noises (babble, general noise, music) from

[1]https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2

MUSAN [Snyder et al., 2015] database, and reverberation that involves convolving room impulse responses (RIR) with audio. Both MUSAN and the RIR datasets are freely available[2]. After data augmentation, utterances under four seconds and speakers with less than eight utterances are removed from the training set. The TDNN is trained to discriminate between speakers. Then, 512-dimensional x-vector speaker embeddings are extracted from layer segment6.

6.2.2.2 Biometric performance evaluation

In table 6.2, we report the biometric performance of the baseline (unprotected) x-vector speaker verification system on the test set of the text-independent VoxCeleb1 database in terms of Equal Error Rate. We report the performance using different back-end scoring and normalization approaches:

System1: As back-end scoring, a classifier based on Probabilistic Linear Discriminant Analysis (PLDA) was trained for the speaker embeddings comparison. Linear discriminant analysis (LDA) was first applied to the speaker's x-vectors extracted from the training set, reducing their dimension from 512 to 200, followed by length normalization and centering using the mean of the training x-vectors. The speakers' x-vectors extracted from VoxCeleb2 (dev and test) and VoxCeleb1 (dev) were used to train the PLDA.

System2: As back-end scoring the cosine distance was used for the speaker embeddings comparison. The x-vectors extracted from the TDNN are passed through the LDA reducing their dimension from 512 to 200 followed by length normalization and centering using the mean of the training x-vectors.

System3: As back-end scoring, the cosine distance was used without applying the LDA and the normalization process to the x-vectors.

As reported in Table 6.2 and Figure 6.5, the biometric performance in terms of EER is equal to 3.12%, 5.5% and 8.18% for system 1, 2 and 3 respectively. For the x-vector-based system trained with softmax loss, the PLDA back-end tends to outperform the cosine since the softmax loss is not discriminative enough to

[2]The data can be download from http://www.openslr.org

Table 6.2: Biometric performance of the the baseline x-vector systems on the test set of VoxCeleb1 text-independent database in terms of EER (%). We report the impact of normalisation and back-end scoring in biometric performance.

Baseline systems	**Normalisation**	**Back-end scoring**	**EER%**
System 1	Mean-centering Length normalisation	LDA+ PLDA	3.12
System 2	Mean-centering length normalisation	LDA + Cosine	5.5
System 3	No normalization	Cosine	8.18

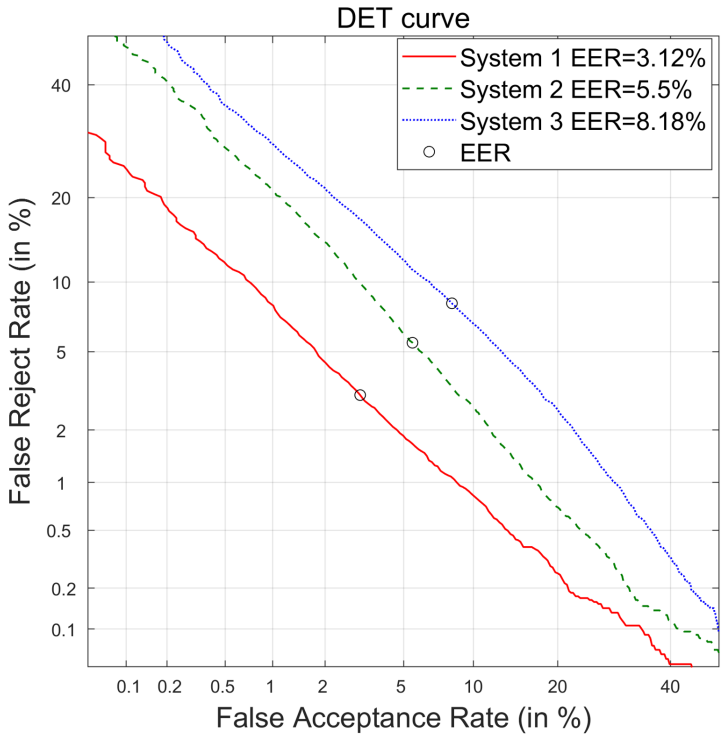


Figure 6.5: DET curves of the baseline x-vectors speaker verification systems, 1, 2, and 3 described in Table 6.2 on the test set of VoxCeleb1 text-independent database.

optimize the embedding similarity. Also, x-vectors contain the lexical content information in the softmax-trained model [Raj et al., 2019]. Thus, the back-end is crucial to deal with the phoneme-invariant problem. For our work, we used the x-vectors extracted using system 3 as a baseline to avoid protecting more biometric

information such as that contained in the PLDA. However, our goal is to develop a privacy-preserving x-vectors system that maintains the biometric performance of the best baseline x-vectors system 1 (PLDA as back-end).

6.3 Binary Representation of x-Vectors Embeddings

Speaker verification systems based on x-vectors use deep neural embedding to represent the speaker. In this section, we present an approach based on deep neural nets autoencoders to extract binary biometric representation from x-vector embedding. This idea was inspired from the binarization of face template [Hmani et al.,] and words vectors [Tissier et al., 2019] where autoencoder architecture was used for transforming real-valued vectors into binary vectors.

6.3.1 Binarization of x-vector based on thresholding method

Before introducing the proposed binarization approach, we first implemented a baseline binarization method based on thresholding. As a baseline x-vector system, we adopt system 3 described in the previous section where 512-dimensional x-vector embeddings are extracted using the TDNN. Then, the speaker's x-vector is binarized using its median by comparing each component to the median value. For the comparison of binary representations, we use the Hamming distance. Table 6.3 reports the biometric performance of the baseline x-vectors system and the binary x-vector system in terms of Equal Error Rate on the test part of text-independent VoxCeleb1 database.

Table 6.3 shows that the method of binarization based on thresholding degrades the biometric performance compared to the baseline x-vectors systems. The EER increases from 8.18% (system 3) to 9.68% with the binary x-vectors. Besides, the dimension of the binary embedding is limited by the dimension of the original x-vector.

Table 6.3: Biometric performance of the speaker verification systems based on the baseline and the binary x-vectors on the test set of VoxCeleb1 text-independent database in terms of EER (%). Binarization is performed with thresholding-method.

Speaker verification systems	**EER%**
Baseline x-vector systems 1/2/3	3.12/5.5/8.18
Binary x-vector with thresholding	9.68

6.3.2 Binarization of x-vectors using deep neural nets autoencoder

In this part, we describe the proposed approach used to transform the real-valued x-vectors embeddings into binary representations based on autoencoder architecture. This approach is based on an autoencoder on top of the TDNN to transform the x-vector embeddings into binary embeddings.

Let $\mathbf{X} = (x_1, ..., x_m)$ be a m-dimensional real-valued vector representing the speaker's x-vector embedding. Our objective is to transform $\mathbf{X}$ into a binary embedding $\mathbf{B} = (b_1, ..., b_n)$ of dimension n independent of the dimension of the original x-vector embedding. For that, the idea was to train an autoencoder model composed of two parts: an encoder that binarizes the x-vector embedding $\mathbf{X}$ to $\mathbf{B}$ and a decoder that reconstructs the x-vector from the binary embedding $\mathbf{B}$.

Encoding to binary embeddings:

In our work, the encoder takes as input the x-vector $\mathbf{X} = (x_1, ..., x_m)$ of a particular speaker extracted using the TDNN model and maps it to a vector $\mathbf{Y} = (y_1, ..., y_n)$ with a dimension equal to the desired binary representation . Then, a binarization layer $\mathbf{B}$ is applied to generate the binary representation $\mathbf{B} = (b_1, ..., b_n)$. The output of the binarization layer is provided by:

$$\mathbf{B}(y_i) = \begin{cases} 0 & \text{if } y_i \leq threshold \\ 1 & otherwise \end{cases} \tag{6.3}$$

Decoding to x-vector embedding

For the decoder, two methods are tested for the training:

- Method1: the decoder is trained to reconstruct from the binary embedding, the speaker's x-vector of a given utterance provided as input to the encoder.
- Method2: The encoder takes as input the speaker's x-vector of given utterance and the decoder is trained to reconstruct from the binary embedding, the average of all the speaker's x-vectors.

For both methods, the autoencoder is trained to minimize the distance between the x-vector embedding reconstructed from the binary embedding and the output given to the decoder.

Figure 6.6 summarizes the binarization approach based on autoencoder model. First, the utterance is fed to the TDNN model to extract a 512-dimensional x-vector speaker embedding. This x-vector is then taken as input to the encoder that maps it to n component real-vector and transforms it to n dimensional binary embedding using the binarization layer. In the end, the decoder reconstructs the x-vector embedding using the binary embedding as a latent representation. This process described in Figure 6.6 (left side) is only used during the training phase. After training, we remove the decoder part, thus an architecture that outputs a binary speaker embedding given a speaker's utterance.

6.3.3 Experimental evaluation and results of binary x-vectors extracted using the autoencoder model

Different configurations are tested in order to find the hyper-parameters (activation function, number of hidden layers, number of neurons, etc.) of the autoencoder that results in the least degradation in biometric performance compared to the baseline x-vector system described in section 6.2.1. The architecture that led to the least degradation was as follows. The encoder consisted of four linear layers with 600, 1000, 1400, and 1000 units respectively using hyperbolic tangent as an activation function. The decoder consisted of two linear layers with 1400 and 1000 units using hyperbolic tangent as an activation function, a linear layer with 600

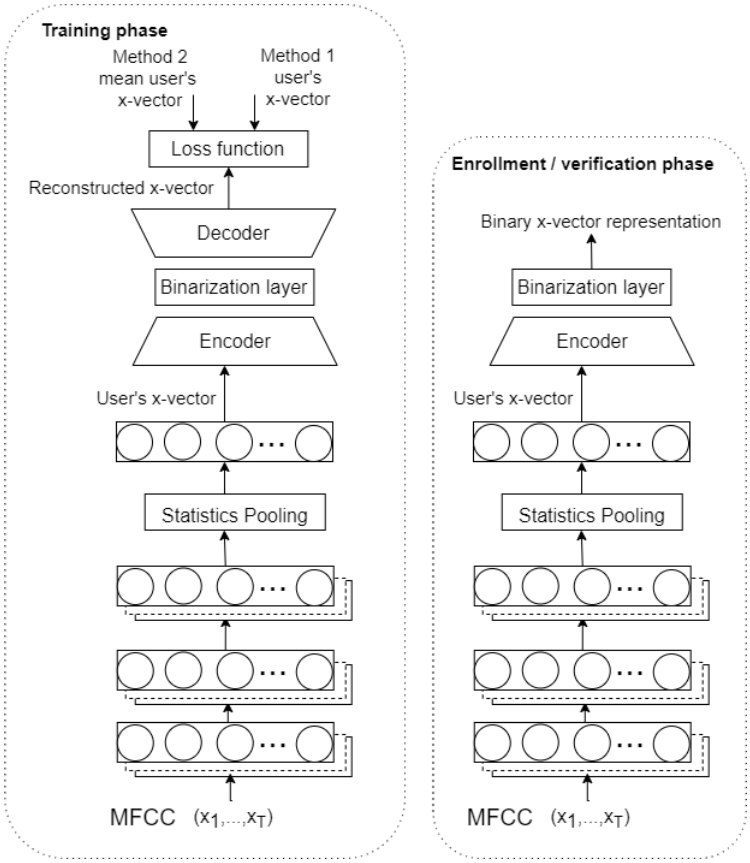


Figure 6.6: Autoencoder architecture used to binarize the x-vector embeddings extracted from the TDNN model. On the left, we show the training phase of the autoencoder. At the right, we remove the decoder part, and as output, we get the binary x-vector representation.

units using ReLU activation function and an output layer with dimension equal to the dimension of the input x-vector with linear activation. The binarization layer is introduced between the encoder and the decoder. The autoencoder training is carried out for 100 epochs in a conventional way by minimizing the Smooth L1 loss between the speaker x-vector reconstructed from the binary embedding and the input x-vector or the average x-vector (method 1 or 2) using Adam optimizer. The learning rate was set to 0.001 with a decay of 0.00001 and the batch size was set to 8000.

The autoencoder was trained using the 512-dimensional x-vectors extracted using the TDNN model of 1 276 888 utterances from 7 323 speakers of the text-

Table 6.4: Biometric performance on the test set of VoxCeleb1 text-independent database of speaker verification system based on binary x-vectors extracted using the autoencoder model. Method 1: decoder trained to reconstruct the x-vector extracted from a given utterance taken as input to the encoder. Method 2: decoder trained to reconstruct the average of speaker's x-vectors.

Binary x-vector dimension	**EER%**	
	Decoder trained with method 1	**Decoder trained with method 2**
800	6.82	3.94
1000	6.64	with binary layer 3.66
		without binary layer 3.87
2000	6.63	4.04
3000	7.45	7.33

independent databases Voxceleb2 [Chung et al., 2018] (dev and test portions) with the training portion (Dev) of VoxCeleb1 [Nagrani et al., 2017]. After training the autoencoder, we remove the decoder part and we obtain as final architecture an encoder at the top of the TDNN that outputs a binary x-vector representation given a speaker's utterance.

Table 6.4 reports the biometric performance of speaker verification system based on binary embeddings. We report the EER on the test set of VoxCeleb1 text-independent database. As shown, the best performance was obtained by using the autoencoder trained to minimize the distance between the x-vector embedding reconstructed from the binary representation and the average x-vector embedding for this speaker. With the autoencoder trained to minimize the distance between the x-vector embedding reconstructed from the binary representation and the x-vector embedding taken as input to the encoder, we notice a degradation in biometric performance.

As shown in Figure 6.7, using method 2 for the training of the autoencoder, the best performance was obtained with a binary embedding of dimension 1000-bits with an EER equal to 3.66%. With the binary embedding of dimensions 800, 2000, and 3000 bits, the performance in terms of EER is equal to 3.94%, 4.04%, and 7.33% respectively. The degradation with 3000-bits could be explained by the lack

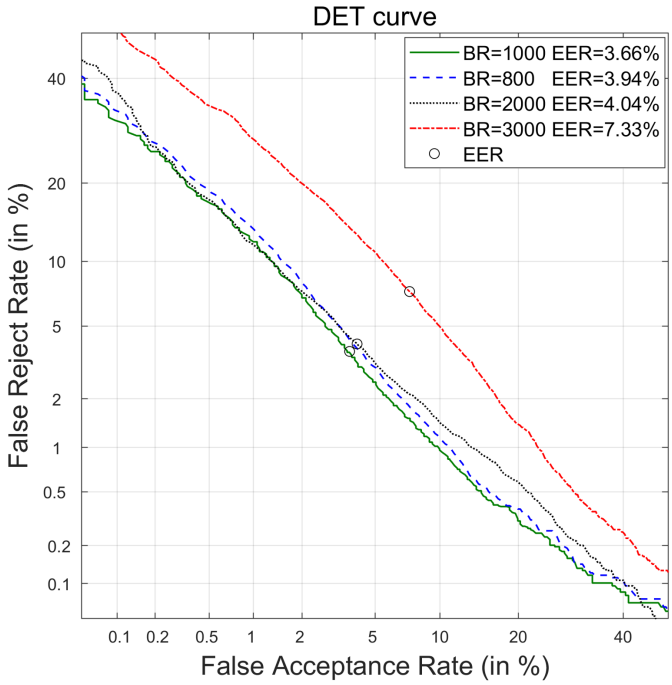


Figure 6.7: DET curves of the speaker verification systems based on binary x-vectors extracted with autoencoder-based method on the test set of VoxCeleb1 text-independent database. We report the DET curves of binary representations (BR) with lengths 800, 1000, 2000, and 3000-bits.

and the loss of information propagated to optimize the parameters of the auto-encoder during the training phase. In fact, by increasing the binary embedding dimension, the number of parameters in the auto-encoder to be optimized becomes bigger.

As shown in Figure 6.8, the performance reported with binary x-vectors is better than the baseline x-vectors systems. Using the x-vectors extracted from system 3 with EER=8.18%, the EER obtained with the binary x-vectors is 3.66%. The performance outperforms that of baseline system 2 based on normalized x-vectors with cosine scoring as back-end (EER=5.5%) and close to the performance of baseline x-vectors system 3 using PLDA as back-end scoring (EER=3.12%).

We have also trained the autoencoder using method 2 without the binarization layer. As reported in Table 6.4, the performance in terms of EER with the x-vectors of dimension 1000 extracted from the last layer of the encoder is 3.87%.

The performance moves from 8.18% using the baseline 512-dimensional x-vectors to 3.87% with the 1000-dimensional x-vectors extracted from the encoder. In fact, by training the auto-encoder to reconstruct the average speaker's x-vectors, we recover the not use of LDA and normalization for the baseline x-vectors.

Using the binarization approach based on the autoencoder model on top of the TDNN, the speaker's x-vector embedding is transformed into binary representation while maintaining the biometric performance. In contrast to the threshold-based binarization method that degrades the biometric performance, binarization of x-vectors using the autoencoder maintains approximately the same performance obtained with the best baseline x-vectors system. In addition, we get control over the length of binary representation by modifying the last hidden layer of the encoder.

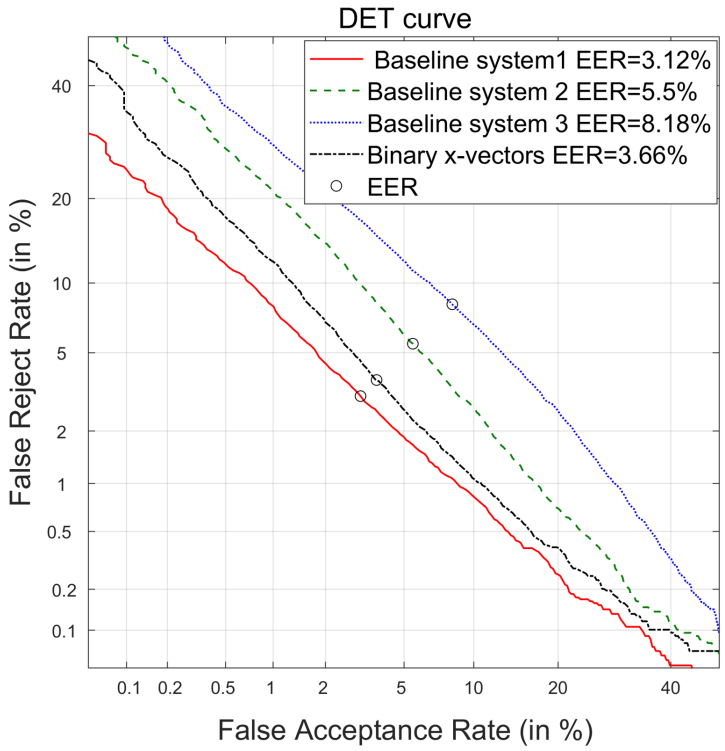


Figure 6.8: Det curves of the baseline and the binary x-vectors speaker verification systems on the test set of VoxCeleb1 text-independent database

6.4 Cancelable x-Vectors

This section analyzes the cancelable x-vectors, which is the output of transforming the binary x-vectors with the shuffling scheme. This transformation allows achieving biometric template protection and privacy requirements. We analyze the distribution of targets and no-targets scores of speaker verification system based on the cancelable x-vectors. This analysis will help us gain insight into the requirements of the error-correcting code and get an idea about the error-correcting capability required to distinguish between intra-domain and inter-domain comparisons.

6.4.1 Experimental evaluation and results of the cancelable x-vectors

After the binarization of x-vector using the autoencoder model described in section 6.3.2, the binary x-vectors (dimension-1000) are transformed with shuffling keys of dimension 1000-bits to generate the cancelable x-vectors. For the evaluation, two scenarios have been considered. Legitimate scenario and stolen key scenarios. For the Legitimate scenario, the impostor (non-target user) does not have information about the shuffling key of the genuine user. The impostor will use his/her biometric data with a random key and tries to access the system. For the stolen key scenario, the impostor has the genuine shuffling key and tries to access the system by presenting the genuine key and his/her impostor's biometric samples.

Cancelable x-vectors are evaluated using the test set of the text-independent VoxCeleb1 database. The target (genuine) and non-target (impostor) scores distributions using the binary and cancelable x-vectors in the legitimate scenario are reported in Figure 6.9. As shown, for the binary x-vectors distributions, there is an overlap between target and no-target distributions. When applying the shuffling scheme, there is a separation of target and non-target scores distributions. In fact, the mean of the target distribution is preserved exactly just like in the binary x-vector level before performing the shuffling transformation. Contrarily, the mean of the non-target scores distribution is augmented when the shuffling scheme is applied and the distribution is right-shifted. As reported in Figure 6.9,

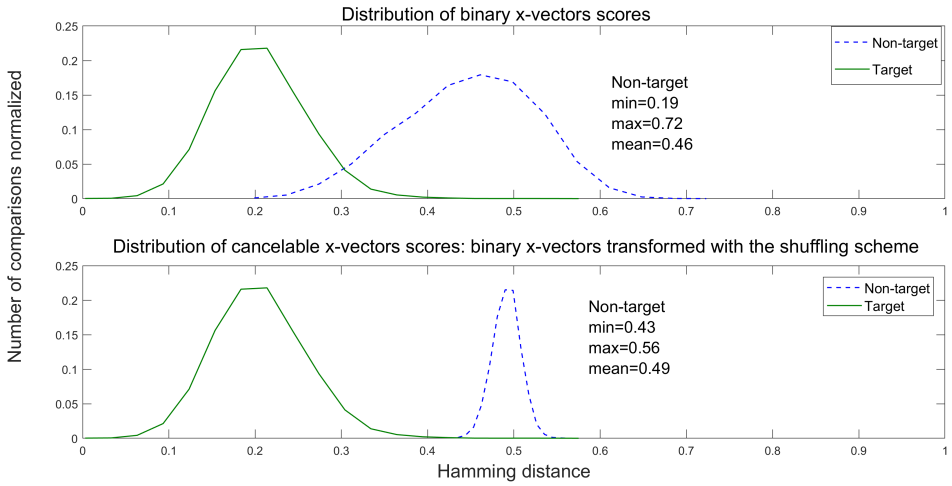


Figure 6.9: Impact of the cancelable shuffling scheme to the binary x vectors: distribution of target and non-target scores for speaker verification systems based on the binary x-vectors and the cancelable x-vectors in the legitimate scenario.

the minimum no-targets score was 0.19 using binary x-vectors and moves to 0.43 when applying the shuffling. Also, the mean of no-targets scores moves from 0.46 to 0.49, which results in no overlapping between target and no-targets distribution using cancelable x-vectors. This leads to an improvement of the biometric performance as shown in the DET curves Figure 6.10. Cancelable x-vector improves the biometric performance with an EER = 0.1% compared to the baseline (PLDA EER=3.12%, cosine EER=5.5%) and the binary x-vectors systems (EER=3.66%).

Regarding the stolen key scenario, we observe in Figure 6.11 that there is an overlap between the stolen key and the target scores distribution. For this scenario, we report in Table 6.5 the False Acceptance Rate obtained according to the FAR and FRR selected in the legitimate scenario.

For the best baseline x-vectors (system1), the EER=3.12%. With the cancelable x-vectors, at the threshold corresponds to FRR=3.12%, the FAR=0 in the legitimate scenario, and the FAR=4.39% in the stolen key scenario (Figure 6.12).

The FAR of the stolen key scenario could be improved by decreasing the threshold taken as a decision for the verification in the legitimate scenario. In fact, as shown in Figure 6.12, by taking the threshold equal to 0.3, the FAR=1.96% in the stolen key scenario. However, by reducing the threshold, even if the FAR in

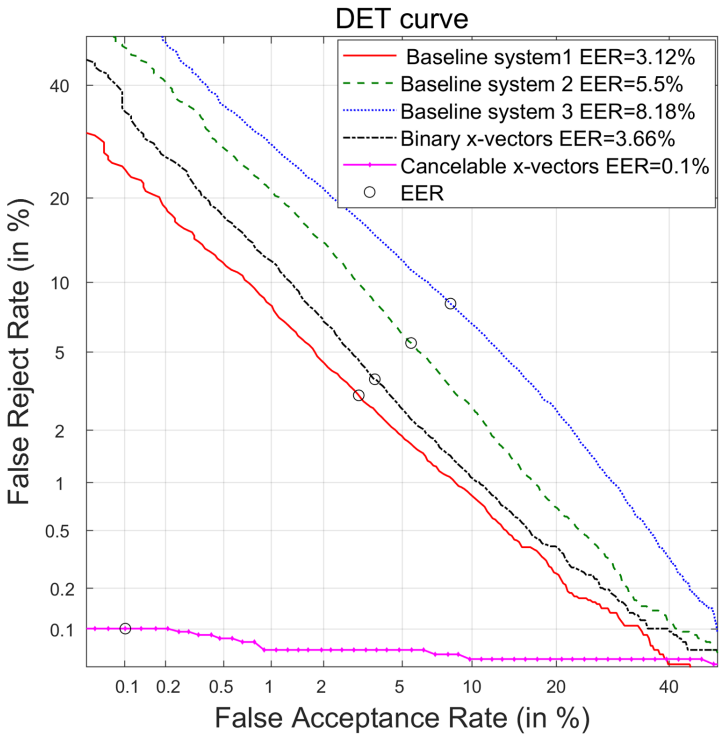


Figure 6.10: DET curves for the speaker verification systems based on the baseline, binary and cancelable x-vectors on test part of VoxCeleb1 text-independent database.

stolen key scenario is improved, the FRR in the legitimate scenario degrades. As an example, for FAR=1.96% in the stolen key scenario, the FRR in legitimate scenario is equal to 7%.

To resolve this issue, the idea was to pass the cancelable x-vector through an error-correcting code to manage the intra-variability which allows to improve the FAR in the stolen key scenario while maintaining the performance in terms of FRR in the legitimate scenario.

Table 6.5: False acceptance rate (FAR) in the stolen key scenario according to the FAR and FRR of the cancelable x-vectors speaker verification system in the legitimate scenario.

FAR-Legitimate scenario	0	0	0	0	0	0
FRR-Legitimate scenario	0.5	1	2	3.04	4	7
FAR-Stolen key scenario	16.15	10.83	6.28	4.39	3.42	1.95

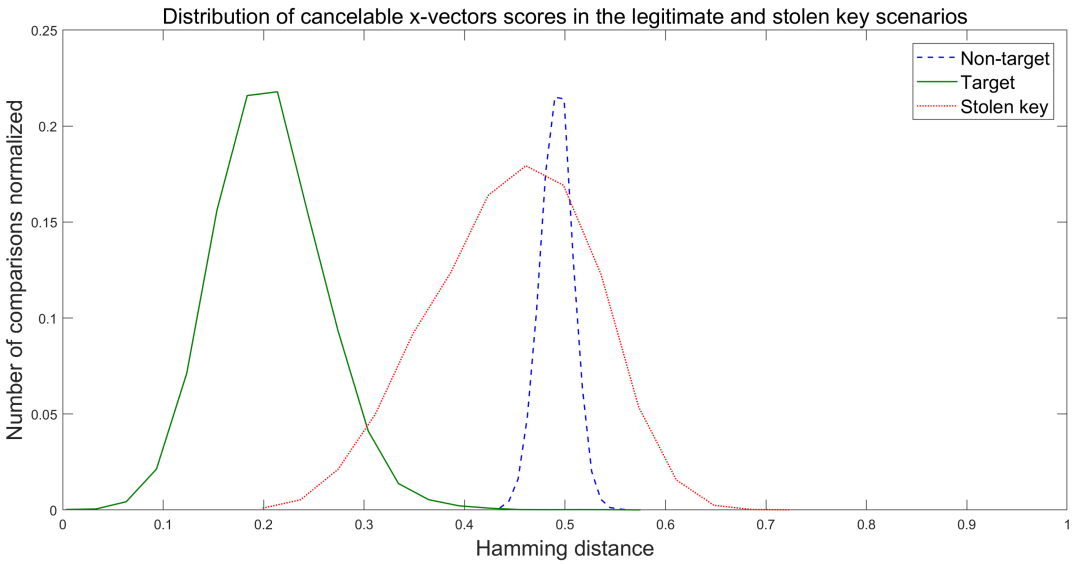


Figure 6.11: Distribution of target, non-target, and stolen key scores for speaker verification system based on cancelable x-vectors.

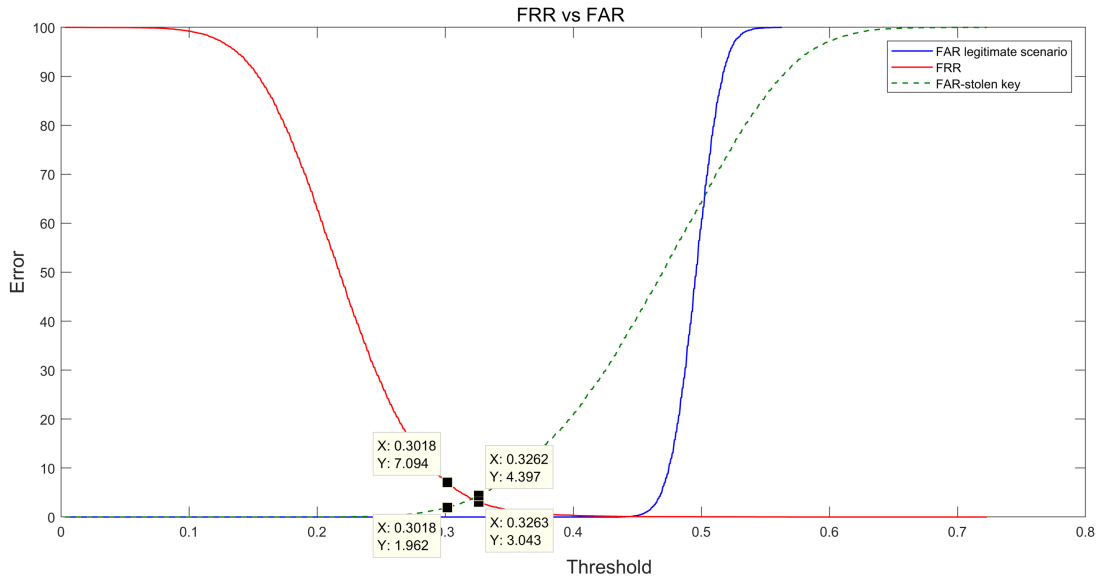


Figure 6.12: FAR and FRR curves of the speaker verification system based on cancelable x-vectors in the legitimate and stolen key scenarios. FRR curve is the same for legitimate and stolen key scenarios since the shuffling transformation preserves the target scores distribution.

6.5 Applying Secure Sketch Error Correction Code to Cancelable x-Vectors

As described in section 6.1, the cancelable x-vectors (output of the transformation of the binary x-vector with the shuffling scheme) is an intermediate template and is not stored in the access system server or the database. The cancelable x-vector is passed through the secure sketch module based on error correcting code to generate a corrected cancelable x-vector which represents the user's template.

6.5.1 Secure sketch error correction code module

For the proposed system, Reed-Solomon (RS) error correction codes [MacWilliams and Sloane, 1977], [Clarke, 2002] was used to extract the secure sketches. It is a linear and cyclic code that belongs to the family of Bose-Chaudhuri-Hocquenghem (BCH) codes. The choice of RS error correction codes was to exploit its maximum distance separable (MDS) property in order to manage the biometric variability. The RS code is described as an $RS(n, k)$ code with m bit-symbols, where n is the block length in symbols and k is the number of information symbols in the message k.

$$k < n \leq 2^m - 1 \tag{6.4}$$

When $n < 2^m - 1$, this is referred to as a shortened form of the RS code. The shortening of the RS code is achieved by making a number of data symbols zero at the encoder, not transmitting them, and then re-inserting them at the decoder. A shortened (n, k) RS code uses (n', k') encoder, where $n' = 2^m - 1$ and $k' = k + (n' - n)$.

By adding $n - k$ parity symbols to the encoded message, RS code can correct t symbol errors, where t is defined as follow:

$$2t = n - k. \tag{6.5}$$

The Reed-Solomon code is also a systematic code, which means that the encoding process does not modify the message symbols k. The codeword n is composed of the original message k appended with the $2t$ parity symbols.

For biometric system, secure sketch [Dodis et al., 2004] consists on the reconstruction of a biometric input w with the help of the secure sketch of w, P and a noisy biometric input w' closer to w. For the proposed system, Reed-Solomon code is used to generate the secure sketch P that corresponds to the $2t$ parity symbols generated by RS encoding of the biometric input w.

Let assume that the binary x-vectors templates are in metric space M with distance function dis, f is the shuffling transformation and w is the binary x-vector. We propose to use RS error correction secure sketch with functions generation Gen and reconstruction Rec and error-correcting capability t as follows. The enrollment function $Enrol$ takes $f(w)$ as input, and outputs the sketch P that corresponds to $2t$ parity symbols:

$$Enrol(w; f) = Gen(f(w)) = P \tag{6.6}$$

For the verification, $verif$ function takes as input the secure sketch P, a probe binary x-vector $w' \in M$, the shuffling transformation f and outputs:

$$verif(w', f, P) = Rec(f(w'), P) \tag{6.7}$$

The correctness property of secure sketches guarantees that if $dis(f(w), f(w'))\leq t$, then $Rec(f(w'), P) = f(w)$. If $dis(f(w), f(w')) > t$, then the reconstruction (decoding) fails and $Rec(f(w'), P) = f(w')$.

For the proposed privacy-preserving x-vectors system, shortened RS code was used. The cancelable x-vector generated by the transformation of the binary embedding with the shuffling scheme is considered as k, the message to be encoded. During step 1 of the enrollment, the cancelable consistent binary x-vector T_c is encoded to extract the $n-k$ parity symbols P which represents the secure sketch. Then during step 2 of enrollment or during authentication, the cancelable x-vector combined with the parity symbols is considered as the noisy codeword, and the RS decoding is performed to reconstruct the corrected cancelable x-vector that corresponds to the template closest to T_c.

6.5.2 Experimental evaluation and results of applying the secure sketch to the cancelable x-vectors

In this section, the output of the whole system is evaluated according to the ISO/IEC 24745 [ISO/IEC JTC1 SC27 Security Techniques, 2011] privacy requirements including biometric performance evaluation, revocability, irreversibility and unlinkability analysis. Moreover, the robustness against different attacks was analyzed.

For the evaluation of the whole system, as a baseline x-vectors system, system3 described in subsection 6.2.2.2 was used where 512-dimensional x-vectors are extracted from the TDNN model. Then, using the binarization method based on autoencoder model integrated on the top of the TDNN, the 512-dimensional x-vectors are transformed into binary embeddings of dimension 1000-bits. Next, the binary embeddings are transformed with shuffling keys of length 1000-bits and passed through the RS error correction code to generate the corrected cancelable x-vectors.

For the proposed system, during step 1 of the enrollment, the consistent cancelable template T_c is segmented into 5×200-dimensional blocks each presenting the message k_c. These blocks are passed through the RS encoder to generate the codewords n, $(n = k_c + 2t)$. At the end of this phase, we store the parity symbols $P = 2t$ corresponding to each block and we delete the rest. The parity symbols present the secure sketch P of this user.

During step 2 of enrollment, the enrollment T_e cancelable x-vector is also segmented into 5 blocks each presenting the message k_e. Then, for each block k_e, the corresponding parity symbols $2t$ (secure sketches) received from the access system are added. The RS decoder takes the couple $(k_e, 2t)$ and performs the decoding process. The result of decoding all the blocks represents the enrollment corrected cancelable x-vector closest to T_c.

During the verification phase, for each block k_p of the probe cancelable x-vector T_p, the secure sketch $P = 2t$ of the claimed identity received from the access system is appended. The RS decoder takes the couple (k_p, P) and assumes that it represents a noisy version of the enrollment template. If $dis(k_c, k_p) \leq t$, then $Rec(k_p, P) = k_c$. If $dis(k_c, k_p) > t$, then the decoding is failing and $Rec(k_p, P) = k_p$.

The result of decoding all the k_p blocks represents the probe corrected cancelable x-vector.

The proposed system was evaluated according to the error-correcting capability of the Reed-Solomon code t. t is the number of errors that the RS code can correct. It is given by $\frac{n-k}{2}$. The error correction capability should be chosen so it is able to distinguish between intra-domain comparisons and inter-domain comparisons. As reported in Figure 6.9 and Figure 6.12 (section 6.4), the verification threshold using only the cancelable x-vectors is 0.44 and the minimum non-target distance is 0.43. Therefore, with cancelable x-vectors of dimension 1000, the error correction capability must be lower than 1000*0.4=400 errors to reduce the intra-variability while preserving the inter variability.

6.5.2.1 Biometric performance evaluation of the cancelable x-vectors after applying the error correction secure sketch

Two scenarios have been considered for the biometric performance evaluation of the proposed system. For the legitimate scenario, the impostor (non-target) will try to impersonate a genuine (target) user by presenting his/her biometrics, a random shuffling key, and the genuine secure sketch received from the access system. For the stolen key scenario, the impostor has the shuffling key of the genuine user and tries to access by presenting his/her biometrics, the genuine shuffling key, and the genuine secure sketch received from the access system.
For step 1 of enrollment, the secure sketch of each speaker is extracted from three random utterances corresponding to this speaker selected from the enrollment utterances on the test set of the VoxCeleb1 database. For the enrollment step 2 and verification phases, we follow the protocol of the VoxCeleb1 test set composed of 18860 target and 18860 nontarget trials.

Legitimate scenario evaluation:

As we mentioned at the end of section 6.4, the idea behind applying the error correction code to the cancelable x-vector was to manage the intra-variability in order to improve the FRR in the legitimate scenario which implies an improvement

in the false acceptance rate for the stolen key scenario.

In table 6.6, we report the biometric performance of the proposed system for the legitimate scenario on the text-independent VoxCeleb1 database test part in terms of FRR and FAR according to different error correction capability t of RS codes. To analyze the impact of t on the biometric performance, the FRR and FAR curves with different error correcting capability t for the legitimate scenario are given in Figure 6.13.

We observe from the reported results that as the error correcting capability t increases, the FRR of the proposed system improves. As reported in Table 6.6, for FAR=0.5% the FRR for $t = 30$, 40, 50, 60 and 70 is equal to 0.079%, 0.074%, 0.058%, 0.047% and 0.047% respectively. This is validated in Figure 6.13, where the FRR is improved by increasing t. As example, at threshold=0.32, the FRR for $t = 30$, 40, 50, 60, and 70 is equal to 3.28%, 2.12%, 1.31%, 0.83%, and 0.31%, respectively. Also, due to the shuffling scheme that separates target and non-target scores distributions, the FAR is close to 0.1% at the EER threshold for the different error correction capability.

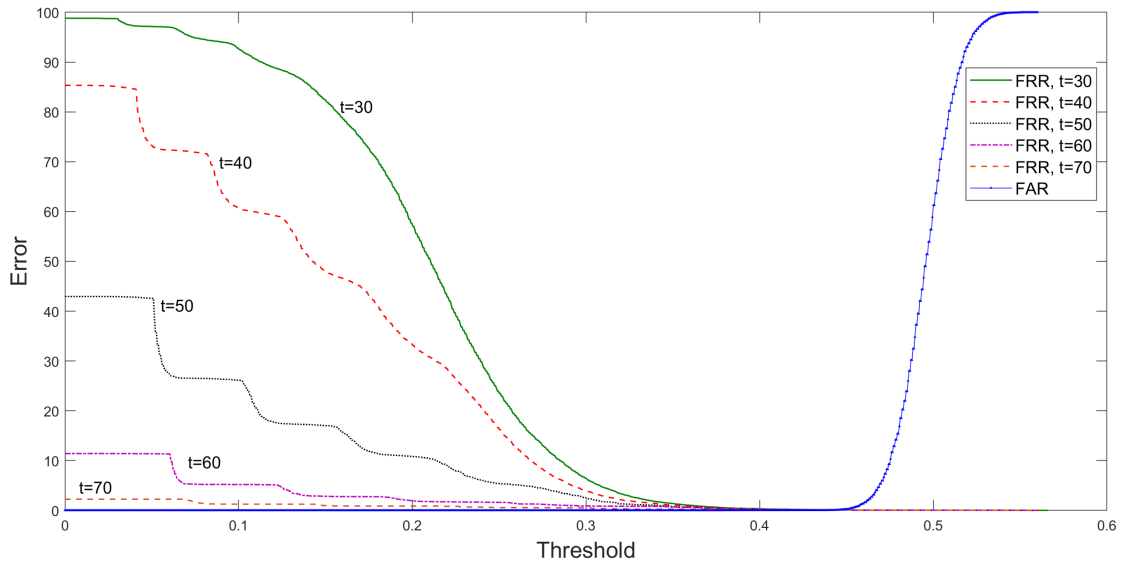


Figure 6.13: FRR curves of the speaker verification system based on the corrected cancelable x-vectors in the legitimate scenario for different error-correcting capability t.

Table 6.6: Biometric performance of the proposed privacy-preserving speaker verification system based on corrected cancelable x-vectors in the legitimate scenario according to the error correction capability of the Reed-Solomon codes. The evaluation is performed on the VoxCeleb1 test part database in terms of FRR and FAR. The cancelable x-vectors is divided into 5 blocks and passed through the RS code, k=200.

RS codes parameters		**Performance in terms of FAR% and FRR% Legitimate scenario**	
n	**t=(n-k)/2**	**FRR**	**FAR**
260	30	0.063	2.8
		0.063	1.18
		0.079	0.57
		0.1	**0.1**
		0.12	0
280	40	0.058	2.78
		0.068	1.15
		0.074	0.51
		0.09	**0.09**
		0.11	0
300	50	0.053	2.73
		0.058	1.11
		0.058	0.53
		0.11	**0.11**
		0.13	0
320	60	0.047	2.67
		0.047	1.17
		0.047	0.56
		0.08	**0.08**
		0.1	0
340	70	0.047	2.7
		0.047	1.19
		0.047	0.58
		0.09	**0.09**
		0.23	0

Figure 6.14 shows the distribution of targets and non-targets scores in the legitimate scenario of the cancelable x-vectors before and after passing through the RS code with t=50. Using these RS parameters, the decoder is able to correct

up to 50 errors for each 200-dimensional block of the cancelable x-vector.

From the distributions of the target scores, we observe that the RS correcting code minimizes the intra-variability. The mean of the target scores distribution moves from 0.22 using only the cancelable x-vector to 0.05 when applying the RS code which explains the improvement in the false rejected rate. However, the mean of non-target scores distribution when applying the RS code is preserved as in the distribution of the cancelable x-vector without RS code. In fact, the RS code cannot correct the inter-variation since the distance between target and non-target cancelable x-vectors is greater than the capability error correction. This leads to the separation of target and non-target distributions, which implies an improvement of the biometric performance. The proposed system outperforms the biometric performance of the baseline x-vectors systems as shown in Figure 6.15. The EER for the proposed system with $t = 30$, 40, 50, 60 and 70 is equal to 0.1%, 0.09%, 0.11%, 0.08% and 0.09%, respectively.

Based on this evaluation, we conclude that the proposed privacy-preserving speaker verification system improves the biometric performance compared to the baseline (unprotected) x-vectors systems based on PLDA or cosine as back-end

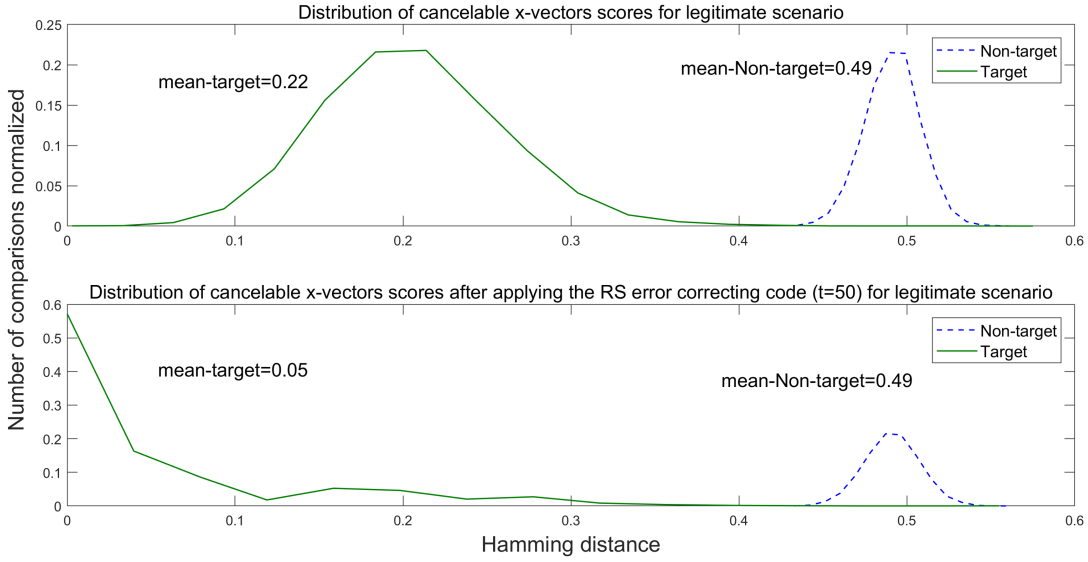


Figure 6.14: Distribution of target and non target scores of the cancelable x-vectors before and after passing through the the Reed-Solomon error correcting code for t=50.

scoring. In addition, as shown in Table 6.7, the proposed system maintains the biometric performance compared to the winners' systems during the VoxCeleb Speaker Recognition Challenge 2019.

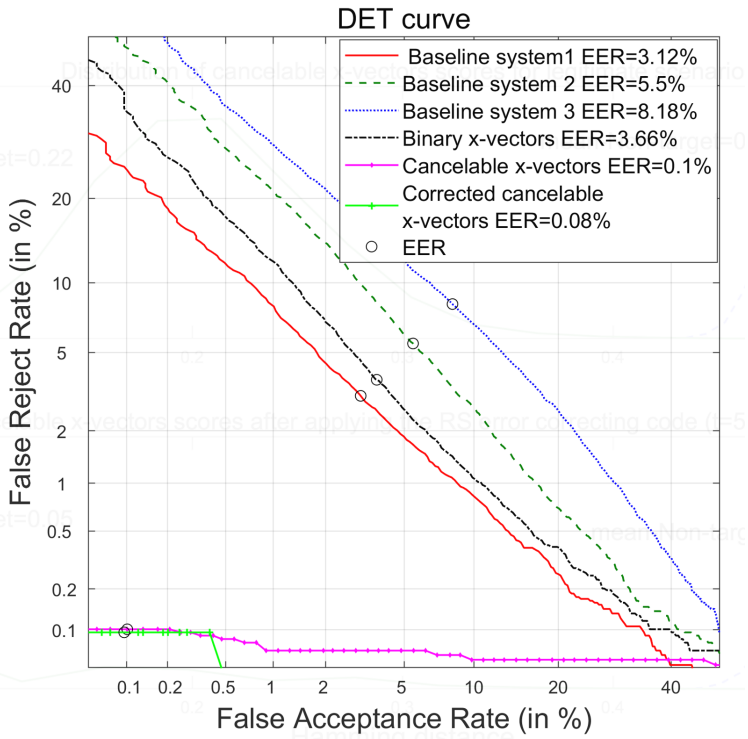


Figure 6.15: DET curves of the speaker verification systems based on the baseline x-vectors, binary x-vectors, cancelable x-vectors and corrected cancelable x-vectors (t=60) on test part of VoxCeleb1 text-independent database.

Table 6.7: Evaluation results on VoxCeleb1 test set for the systems submitted during the VoxCeleb Speaker Recognition Challenge 2019.

System	BUT [Zeinali et al., 2019]		JHU-HLTCOE [Garcia et al., 2020]		[Zhou et al., 2019]	**Proposed cancelable x-vectors**
	Single system	Fusion 4 systems	Single system	Fusion 4 systems	Single system	
EER%	1.22	0.96	1.74	1.54	1.85	**0.1**

Stolen key scenario evaluation:

For the stolen key scenario, we report in Table 6.8 the FAR according to the FAR and FRR of the proposed system in the legitimate scenario. For the best unprotected x-vector system based on PLDA as back-end scoring the EER=3.12% (FAR=FRR=3.12%). For the proposed system, using RS code with t=50, at the decision threshold corresponding to FRR= 1.01%, 2.08%, 3.01%, and 3.38%, the FAR=0 in the legitimate scenario and the FAR on the stolen key scenario is 7.2%, 2.79%, 2.06% and 1.8% respectively. A clear improvement is shown in terms of FAR when comparing these results with that reported in Table 6.5, where we reported the FAR of the stolen key scenario using the cancelable x-vector without applying the Reed-Solomon error correction code. For example, at the threshold corresponding to FRR=%2 and FAR=0 in the legitimate scenario, we get for stolen key scenario a FAR=6.28% using only the cancelable x-vectors and FAR=2.79% using the corrected cancelable x-vectors.

For the analysis of the impact of error correction capability t on the FAR in the stolen key scenario, we reported in Table 6.9 the FAR obtained considering FRR=2% and FAR=0 in the legitimate scenario according to t. We can observe that for $t = 30$ up to $t = 50$, as t increases, the FAR of the stolen key improves and the best value is obtained for $t = 50$ with FAR = 2.79%. Then for $t = 60$ and $t = 70$, we observe that the FAR degrades compared to the FAR obtained at $t = 50$.

In fact, for the stolen key scenario, the same shuffling key is used to transform the binary x-vectors of the target and non-target users. In this case, the distribution obtained for the cancelable x-vectors will be the same as the distribution of the binary x-vectors before the transformation. For this, the RS code with error correction capability tuned according to the target and non-target score distributions of cancelable x-vectors in the legitimate scenario will now be applied to the score distributions of binary x-vectors.

For the score distributions of cancelable x-vectors in the legitimate scenario (Figure 6.9, section 6.4), there is no overlap between target and non-target distributions. In this scenario, the RS code will reduce the intra-variability, but it cannot correct the inter-variability since the distance between the target and

non-target templates is greater than the capability error correction. However, the distribution of the cancelable x-vectors in the stolen key scenario is the same distribution obtained with the binary x-vectors, applying a slight overlap between target

Table 6.8: Biometric performance in terms of FAR for the proposed privacy-preserving speaker verification system based on corrected cancelable x-vectors in the stolen key scenario. The evaluation is performed on the VoxCeleb1 test part according to different Reed-Solomon codes. The cancelable x-vectors is divided into 5 blocks and passed through the RS code, k=200.

RS codes parameters		**Biometric performance of the proposed system at the stolen key scenario in terms of FAR%**		
n	**t=(n-k)/2**	**Legitimate scenario**		**Stolen key scenario**
		FRR	**FAR**	**FAR (@ FAR and FRR of legitimate sceario)**
260	30	1.03	0	10.49
		2.06	0	5.67
		3.02	0	3.96
280	40	1	0	9.21
		2	0	4.17
		3	0	2.4
300	50	1.01	0	7.2
		2.08	0	2.79
		3.01	0	2.06
		3.12	0	1.94
320	60	1	0	6.6
		2.06	0	3
		3.04	0	2.2
340	70	1.03	0	7.9
		2.01	0	4.45
		2.24	0	3.9

Table 6.9: FAR in the stolen key scenario according to the error correction capability t for speaker verification system based on the corrected cancelable x-vectors. The FAR is reported at FRR=2% and FAR=0 in the legitimate scenario.

Error correcting capability t (bits)	30	40	50	60	70
FAR-stolen key %	5.67	4.17	2.79	3	4.45

and stolen key distributions. Therefore, when applying the RS code to reduce the intra-variability, it will also correct the inter-variability and a clear correction will be observed when we take an RS code with a high error-correcting capability t.

The distributions of target and stolen key scores of cancelable x-vectors before passing through the RS code and after applying the RS code using different t are given in Figure 6.16.
For t=30, the distribution of cancelable x-vectors after applying the secure sketch is close to the distribution of cancelable x-vectors without secure sketch and the FAR in the stolen key scenario is equal to 5.67% (at FRR=2% for legitimate).

For t=50, the intra-variability is reduced, the mean of target scores moves from 0.22 to 0.05 while preserving the mean of the distribution of stolen-key scores. In fact, the EER threshold between target and stolen key distributions is 0.32 meaning 320 mismatch bits with cancelable x-vectors of 1000-bits. When t = 50, the RS code can correct up to 250 errors which are in the range of target scores. Therefore, only the mean of target scores is reduced. As result, the overlapping between target and stolen key distributions is decreased which improves the FAR

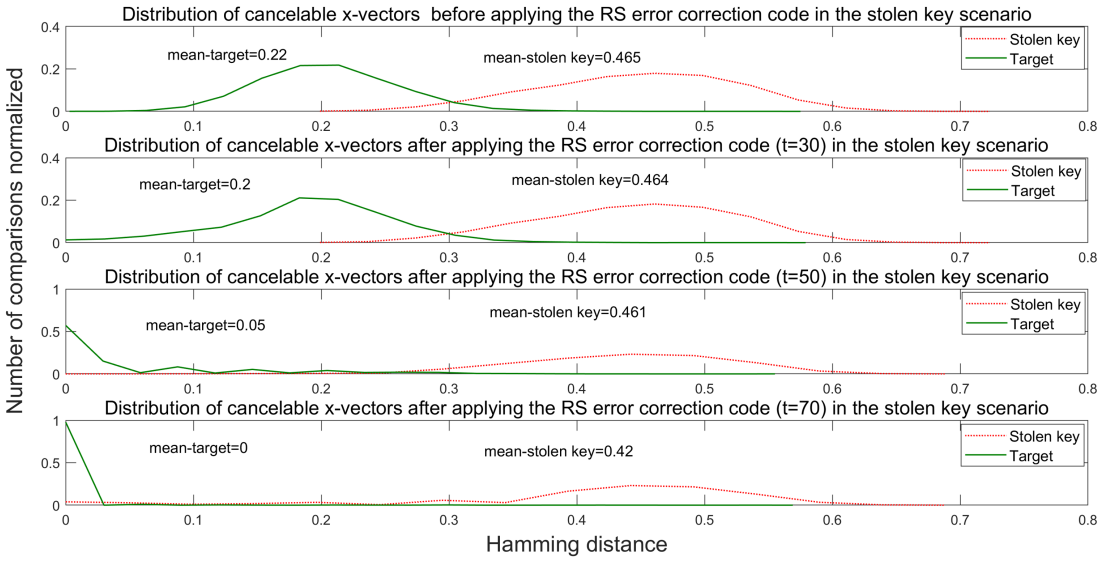


Figure 6.16: Distribution of target and stolen key scores for the privacy-preserving speaker verification system based on corrected cancelable x-vectors according to different error correcting capability.

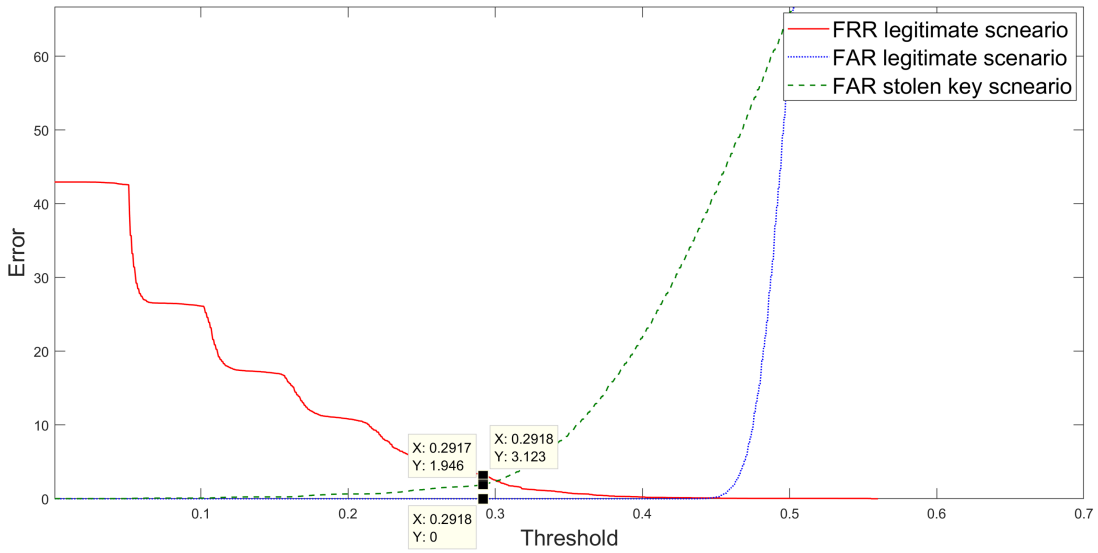


Figure 6.17: FRR and FAR curves of the speaker verification system based on the corrected cancelable x-vectors in the legitimate and the stolen key scenarios using $t = 50$.

at stolen key scenario to 2.79% (at FRR=2% for legitimate).

For t=70, the intra-variability is reduced and the mean of target scores moves from 0.22 to 0. However, the mean of stolen key distribution is also reduced and moves from 0.46 to 0.42 because with t=70, the RS code can correct up to 350 errors which are in the range of stolen key distances. This implies overlapping between targets and stolen key distribution which explains the increase of the FAR with t=70 to 4.45% compared to FAR=2.79% with t=50 as reported in Table 6.9.

For the baseline x-vectors system, the best biometric performance using PLDA as back-end scoring in terms of EER is 3.12% (FAR=FRR=3.12%). For the proposed privacy-preserving speaker verification system based on corrected cancelable x-vectors, as shown in Figure 6.17, at the threshold corresponds to FRR=3.12%, the FAR = 0 in the legitimate scenario, and the FAR=1.94% in the stolen key scenario. The system outperforms the performance of the baseline system even in the stolen key scenario.

As shown in Figure 6.18, by first applying the shuffling scheme, we separate the target and non-target distributions, which allows to improve the performance in terms of false acceptance rate. Then, due to the extraction of the secure sketch

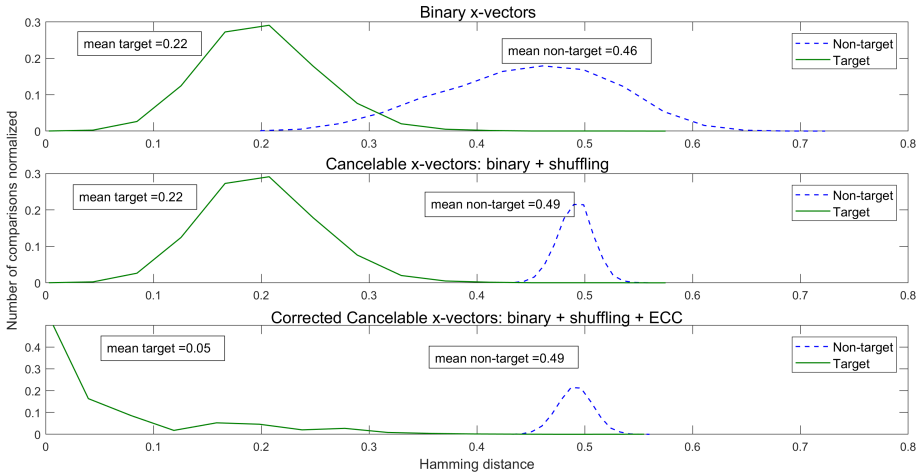


Figure 6.18: Distribution of target and non target scores of the binary, cancelable and corrected cancelable x-vectors.

from the three enrollment utterances (consistent x-vector), we can correct the cancelable enrollment and probe x-vectors to be close or equal to the cancelable consistent x-vector. This allows to reduce the intra-variability which improves the performance in terms of false rejection rate.

For the proposed system, the enrollment phase requires four utterances for each speaker. During step1, three utterances are used to extract the secure sketch. Then during step 2 of enrollment, the corrected cancelable enrollment template is generated using the fourth utterance and the secure sketch extracted during step1. This process makes it possible to correct the cancelable enrollment template and to extract a secure sketch per speaker and not per utterance.

The proposed system can be adapted to perform with one utterance instead of four during the enrollment phase as shown in Figure 6.19. In this case, during the enrollment phase, the user provides its enrollment voice sample to extract the binary x-vector using the TDNN and the auto-encoder models. Then, the binary x-vector is transformed using the user-specific shuffling key K_e to generate the enrollment cancelable x-vector T_e. Next, T_e is passed through the RS encoder to extract the secure sketch P which corresponds to the parity symbols extracted from the RS encoding of T_e. At the end of enrollment phase, we store the secure sketch P and the enrollment cancelable x-vector T_e which represents the user's

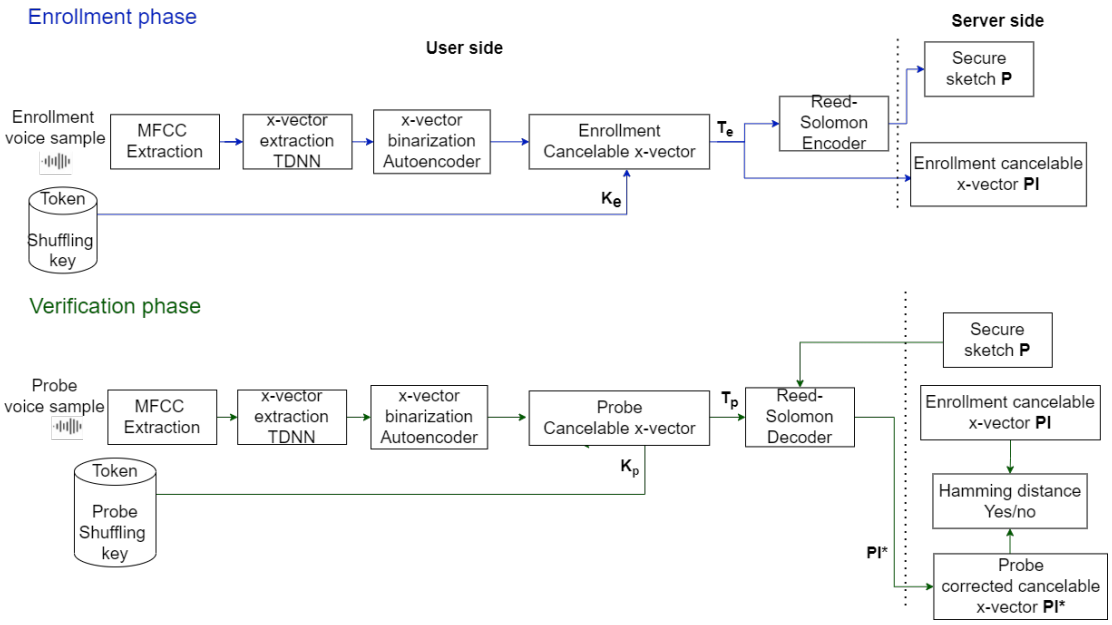


Figure 6.19: Pipeline of enrollment and verification phases for the privacy-preserving speaker verification system based on one utterance during the enrollment phase.

enrollment template PI on the access control system and we delete the rest. During the verification phase, the user presents the probe voice sample and the probe shuffling key. The probe sample is passed through the x-vector embedding extractor and binarization modules to obtain the probe binary x-vector. Using the shuffling key provided by the user, the probe cancelable x-vector T_p is generated. Then, the RS decoder assumes that T_p is an error-prone version of T_e, it combines the secure sketch P of the claimed identity with T_p, and performs the decoding process to generate PI^* which represents the user's probe corrected cancelable template.

In contrast to the system based on four utterances for the enrollment, with this process, the enrollment cancelable x-vector T_e is stored without applying the error correction. Only the probe cancelable x-vector T_p is corrected to be close to T_e.

In Table 6.10, we report the biometric performance of the adapted system for the legitimate scenario on the test set of VoxCeleb1 database. Even with one utterance for the enrollment phase, the speaker verification system based on corrected cancelable x-vectors improves the performance in the legitimate scenario

Table 6.10: Biometric performance in the legitimate scenario for the system based on corrected cancelable x-vectors using one utterance for the enrollment phase. The evaluation is performed on the VoxCeleb1 test set with error correction capability t=50.

Performance in terms of EER%	**Baseline x-vectors**			**Binary x-vectors**	**Corrected cancelable x-vectors**
	LDA + PLDA	**LDA + Cosine**	**Cosine**		
EER%	3.12	5.5	8.18	3.66	0.1

Table 6.11: Biometric performance in terms of FAR for the privacy-preserving speaker verification system based on corrected cancelable x-vectors in the stolen key scenario.

Systems	**Scenario**		**FAR% @ FRR=3.12%**
Baseline x-vectors	Legitimate		3.12
Corrected cancelable x-vectors	Legitimate		0
	Stolen key	Four enrollment utterances	1.94
		One enrollment utterance	4.1

with EER=0.1% compared to the baseline x-vectors with PLDA scoring as back-end (EER=3.12%). For the stolen key scenario, as reported in Table 6.11, at the threshold corresponds to FRR=3.12%, a FAR=4.1% was reported. We observe degradation in terms of false acceptance rate for the stolen key scenario compared to the system based on four utterances for the enrollment phase. The FAR moves from 1.94% for the system based on four enrollment utterances to 4.1% for the system based on one utterance. Therefore, in case a robust privacy-preserving speaker verification system is required, the proposed system based on four utterances of enrollment could be a solution since it maintains the biometric performance in the legitimate and the stolen key scenarios. On the other hand, in the case where the robustness against the stolen key scenario is not a priority or in the case of the unavailability of several enrollment utterances, the system based on one utterance during the enrollment could be used.

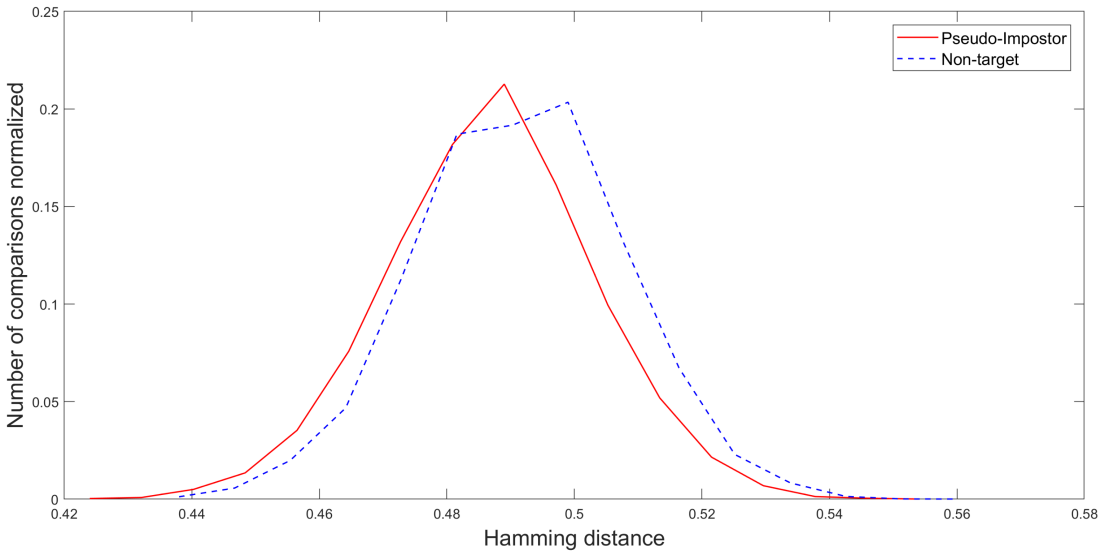


Figure 6.20: Revocability analysis: Distribution of Non-target and pseudo-impostor scores using VoxCeleb1 test set.

6.5.2.2 Revocability analysis

The proposed privacy-preserving x-vector speaker verification system achieves the revocability requirements. In case the corrected cancelable x-vector is compromised, a new one can be generated using the same biometric sample by changing the shuffling key. For the proposed system, the secure sketch is extracted from the RS encoding of the cancelable x-vector. Therefore, by changing the shuffling key, we can generate a new cancelable template which results in a new secure sketch and a new corrected cancelable x-vector.

Revocability is evaluated by computing the pseudo-impostor scores. The pseudo-impostor is the comparison of a corrected cancelable x-vector of a particular user generated from a biometric sample X with new templates generated using the same X and different shuffling keys. We have performed 100,000 comparisons for each user in the test part of Voxceleb1 (40 users). As shown in Figure 6.20, the distribution of the pseudo-impostor scores overlaps with the non-target distribution which means that the new generated cancelable templates are indistinguishable from each other, although they are generated from the same voice sample. As a result, in case of compromise, a cancellation is possible, and a new template can

be generated from the same voice sample by changing the shuffling key.

We estimate the number of possible new templates using the Hamming bound [MacWilliams and Sloane, 1977] as described in the evaluation of revocability of protected i-vector (Chapter 5, section 5.3.4). For the proposed privacy-preserving x-vector system using Reed-Solomon code with error capacity equal to 60, the threshold at $EER = 0.08\%$ is $th = 0.44$ and the probability of bits equal to 1 is 0.45 for cancelable x-vectors of dimension 1000-bits. Therefore, we get almost 2^{28} possible new cancelable templates PI for each user as given by Eq 6.8.

$$\begin{aligned} PI &= \frac{Number\ of\ possible\ permutation}{Volume\ of\ Hamming\ spheres} \\ &= \frac{1000!}{(450!)(550!)\sum_{k=0}^{(th\times 1000)}\binom{1000}{k}} \approx 2^{28} \end{aligned} \tag{6.8}$$

6.5.2.3 Unlinkability analysis

As defined in [ISO/IEC JTC1 SC27 Security Techniques, 2011], the cancelable x-vectors generated from the same biometric samples should not be linkable across databases and applications. The goal of this evaluation is to determine if from two corrected cancelable x-vectors T1 and T2 enrolled in different applications, we can know whether they are generated from the same user or not.

For the unlinkability analysis, we use the framework defined in [Gomez-Barrero et al., 2017]. This protocol is based on Mated and Non-Mated distributions. *Mated instances:* scores computed by comparing corrected cancelable x-vectors extracted from different samples of the same subject using different shuffling keys. *Non-mated instances:* scores computed by comparing corrected cancelable x-vectors generated from samples of different subjects using different shuffling keys. For an unlinkable system, we should have an overlap between the mated and non-mated distributions.

As described in [Gomez-Barrero et al., 2017], the global metric $D_{\leftrightarrow}^{sys}$ gives an estimation of the global linkability of the system. If a system has $D_{\leftrightarrow}^{sys} = 1$, where both score distributions (mated and non-mated) have no overlap means that the system is fully linkable. If a system has $D_{\leftrightarrow}^{sys} = 0$, where both score distributions

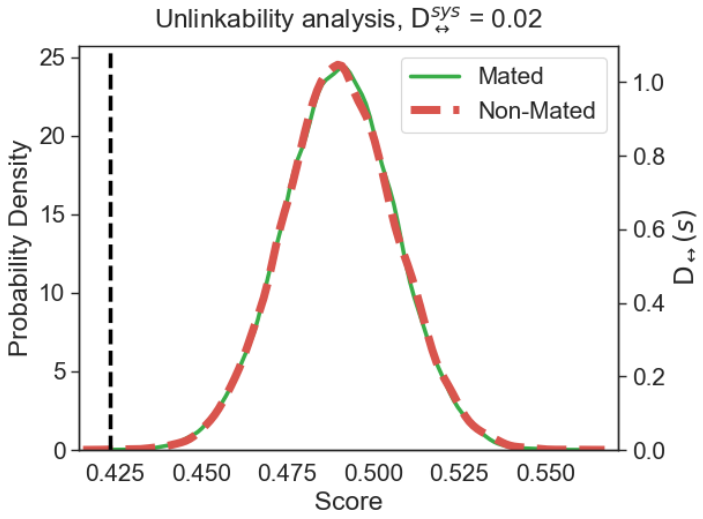


Figure 6.21: Unlinkability analysis: Distribution of Mated and Non-Mated scores using VoxCeleb1 test set.

are totally overlapped means that the system is fully unlinkable for the whole score range. As observed in Figure. 6.21, the distribution of mated and non-mated scores are overlapped with global linkability $D_{\leftrightarrow}^{sys}$ close to 0 ($D_{\leftrightarrow}^{sys} = 0.02$). Based on this evaluation, the proposed system is considered unlinkable.

6.5.2.4 Irreversibility analysis

This subsection analyzes the privacy leakage of the user's biometric information for the proposed privacy-preserving x-vectors system. We will suppose that privacy is compromised if the attacker succeeds to reconstruct the user's binary embedding B. The information leaked can be presented as mutual information:

$$I(B, E) = H(B) - H(B|E) \tag{6.9}$$

where B represents the user's binary embedding and E represents the information that an attacker can compromise. For our system, E could be the user shuffling key k and/or the user's corrected cancelable x-vector PI. $H(B)$ represents the entropy of B and computes the number of bits required to specify B. $H(B|E)$ is the entropy of B given E.

Table 6.12 reports the entropy [Shannon, 1948] of the binary embeddings according to their length. The entropy is measured on 4874 samples of the test part of VoxCeleb1. Binary embeddings with lengths 800 and 1000-bits provide the highest entropy with 756 and 936 respectively. Binary embedding with a length of 3000 bits gives the lowest entropy value, which is correlated with their biometric performance.

Table 6.12: Entropy of the binary embeddings extracted using a decoder trained to reconstruct the average of speaker's x-vectors. The entropy was measured using 4874 samples of the test set of VoxCeleb1 database.

Binary embedding length	**EER%**	**Entropy**
800	3.94	756
1000	3.66	936
2000	4.04	1788
3000	7.33	1809

The irreversibility of the proposed system is evaluated under different attack scenarios:

1) Shuffling key k is compromised: In this scenario, we suppose that the attacker gain access to the user's shuffling key. In this case, $E = k$ and the mutual information is given by:

$$I(B, k) = H(B) - H(B|k) = 0 \tag{6.10}$$

$I(B, k) = 0$ because $H(B) = H(B|k)$ as the user's shuffling key does not provide any information about the user's binary embedding. The shuffling key k only provides the positions of bits after the transformation of B but does not gives the values of bits in the binary embeddings.

2) Corrected cancelable x-vector PI is compromised: In this scenario, the attacker gain access to the user's corrected cancelable x-vector and the secure sketch stored on the server-side. We consider the worst-case scenario and we assume that the attacker succeeds to recover the user's cancelable x-vector T. In this case,

$E = T$ and the mutual information is given by:

$$I(B, T) = H(B) - H(B|T) \tag{6.11}$$

where $H(B|T)$ measure the unpredictability of B given T. Without having information about the user shuffling key, the attacker does not know the exact locations of the bits in the binary embeddings. In fact, without knowing the shuffling key and prior knowledge about the distribution of the binary embeddings, it is computationally not feasible to revert to the original binary embeddings B as the number of permutations to be tested is too big. For the proposed system, if the attacker wants to guess the correct positions of the binary embedding of length 1000-bits and knowing the probability of bits equal to 1 is 0.45, the guessing complexity is equal to 2^{994} the number of possible permutations, given by Eq. 6.12 as follows:

$$Number\ of\ possible\ permutations = \frac{1000!}{(450!)(550!)} \approx 2^{994} \tag{6.12}$$

Also, as reported in the unlikability analysis, the protected cancelable x-vectors are unlinkable, which means that the cancelable x-vector T and the binary embedding are independent. Therefore, the reconstruction of B given T is not possible.

As explained above, it is difficult to recover the binary x-vector when the

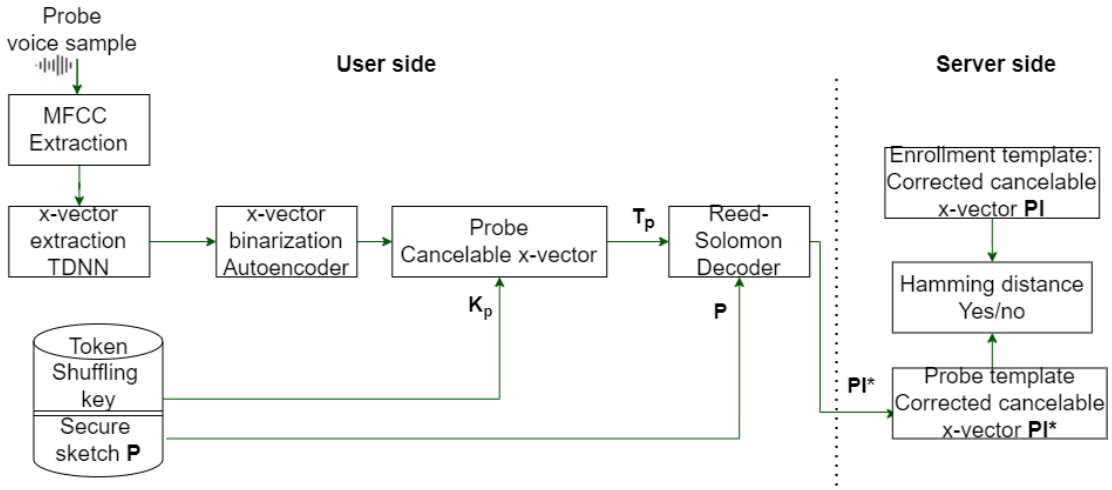


Figure 6.22: Pipeline for the verification phase when the secure sketch is stored on the client-side.

corrected cancelable x-vector and the secure sketch stored on the server-side are compromised. However, storing the secure sketches on the server side could impact the privacy of the users. In case the server is attacked, the secure sketches of all the users will be compromised and become public. Therefore, to improve the proposed system, it's better to store the secure sketch on the client-side as shown in Figure 6.22. When a client's token is compromised, it will contain only his secure sketch.

3) Corrected cancelable x-vector PI and shuffling key are compromised: In the case the attacker gain access to both cancelable x-vector T and the shuffling key k, the attacker can reconstruct the original binary embedding because knowing T and k provide the information about the values and the positions of bits in the binary embedding.

6.5.2.5 Security analysis

In this section, we evaluate the security of the proposed privacy-preserving speaker verification system based on x-vectors against different scenarios of attacks. We compute the false acceptance rate for each attack scenario when the EER threshold of the proposed system ε_{EER} is taken as the decision threshold. A high value of FAR implies that the system is not robust to this attack scenario.

Stolen biometric attack:
In this scenario, we suppose that the attacker gain access to the target user biometric sample. Then, he/she tries to impersonate the target user by presenting the stolen biometric sample, a random shuffling key and the target secure sketch received from the access system.

The corrected cancelable x-vector is the output of the RS decoding of the cancelable x-vector combined with the target user secure sketch. Moreover, the cancelable x-vector is generated by the transformation of the binary embedding (extracted from the biometric sample) with the user-specific shuffling key. For our system, even if the attacker presents the biometric sample of the target user, the transformation with random shuffling key results in a separation between the distribution of the scores corresponding to the cancelable x-vectors of the attacker

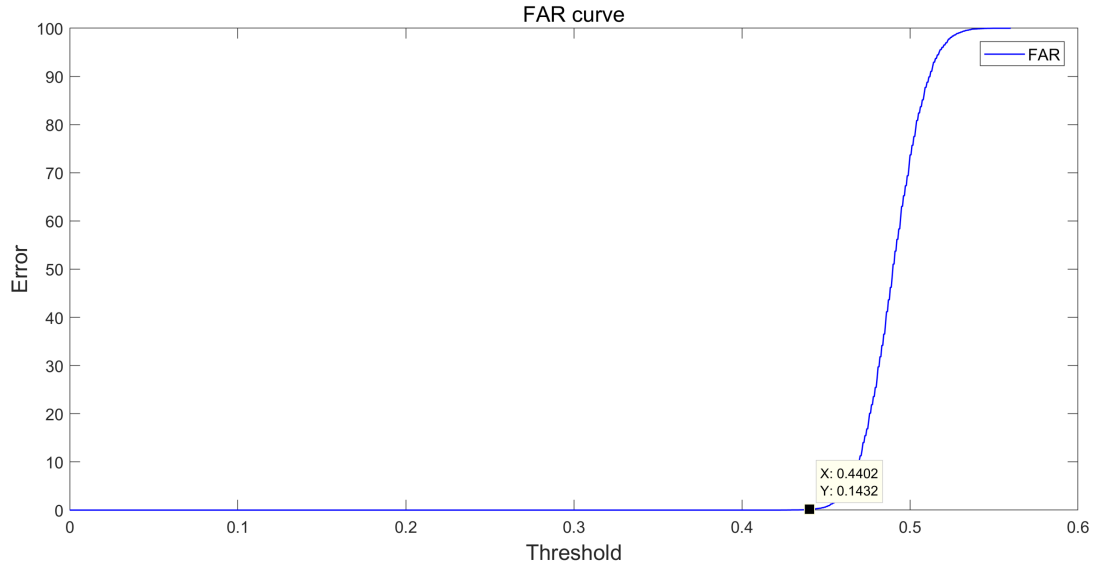


Figure 6.23: FAR curve of the proposed privacy-preserving speaker verification system in the stolen biometric attack scenario using VoxCeleb1 test set. FAR=0.14 at the EER threshold of the legitimate scenario.

and that of the target user. This implies that the RS decoding can not correct the attacker's cancelable x-vector since the distance is greater than the capability error correction.

As shown in Figure 6.23, the FAR obtained when the attacker gains access to the target biometric sample is equal to 0.14% at the EER threshold of the legitimate scenario. Based on this evaluation, we conclude that the proposed system is robust to stolen biometric attacks.

Brute force attack

The decision on the proposed speaker verification system is based on comparing the Hamming distance between the enrollment and the probe corrected cancelable x-vectors to a specific threshold. If the distance is less than this threshold, the user is considered a legitimate user. The brute force attack consists of an attacker trying to guess the target-enrollment template's values to gain access. However, this is infeasible in our system because the possible combinations are huge. For the proposed system, the dimension of the template is 1000 bits and using a Reed Solomon with t= 60, the access threshold for verification is 0.4 making the guess-

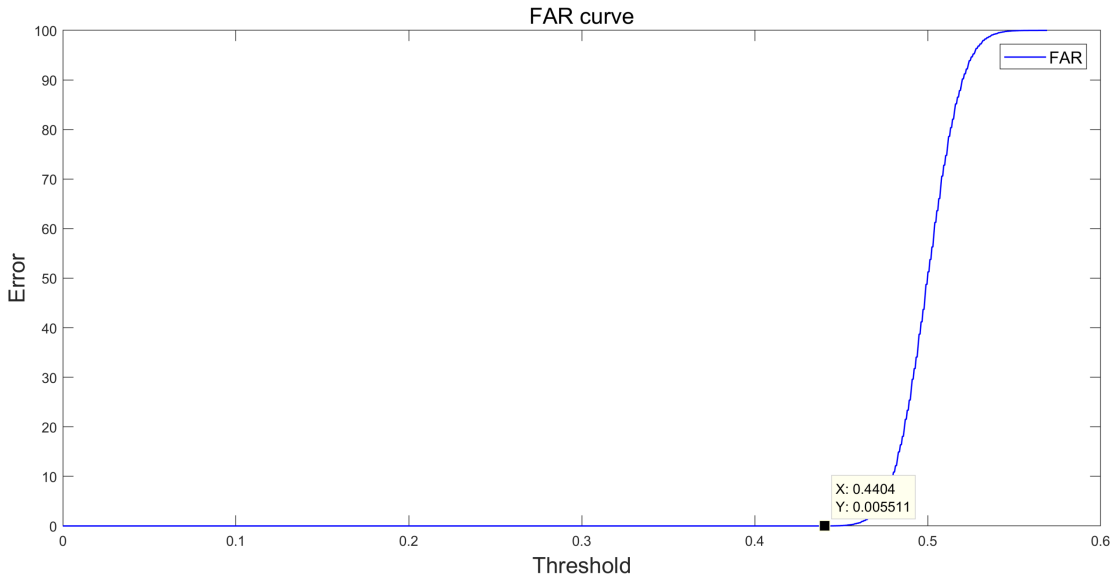


Figure 6.24: FAR curve of the proposed privacy-preserving speaker verification system in the brute force attack scenario using VoxCeleb1 test set. FAR=0 at the EER threshold.

ing complexity equal to $2^{1000*(1-0.4)}$ attempts.

For the evaluation of the brute force attack, we attacked the corrected cancelable x-vector of each user in the test part of Voxceleb1 (40 users) with 100,000 synthesized templates. As shown in Figure 6.24 the FAR=0 for this attack at the EER threshold.

Stolen token attack

In this scenario, the attacker has stolen the shuffling key of the target user and tries using random binary vectors to generate the taregt's cancelable x-vector. For the evaluation of this attack scenario, we attacked the corrected cancelable x-vector of each user in the test part of Voxceleb1 (40 users) with 100,000 cancelable x-vectors generated from 100,000 random binary vectors transformed with the target user shuffling key. As shown in Figure 6.25 the FAR=0 for this attack at the EER threshold of the final system.

Worst case scenario:

The worst-case scenario corresponds to the stolen key scenario evaluated in sec-

tion 6.5.2.1. The attacker has the shuffling key of the user and tries to access the system by presenting his/her biometrics, the target shuffling key, and the target secure sketch received from the access system. One of the main requirements in the standard ISO/IEC 24745 for biometric information protection is that the protection of biometric systems should not degrade the biometric performance compared to the baseline system. In our case, if we consider the best baseline x-vectors system based on PLDA as back-end scoring, the biometric performance in terms of EER=3.12% (FAR=FRR=3.12%). As shown in Figure 6.26, for the proposed privacy-preserving speaker verification system based on cancelable x-vectors, at FRR=3.12%, the FAR=0 in the legitimate scenario and the FAR=1.94% for worst case attack (stolen key). The proposed system outperforms the performance of baseline system in legitimate and stolen key scenarios. Therefore, we conclude that the proposed system is robust against stolen shuffling key scenario.

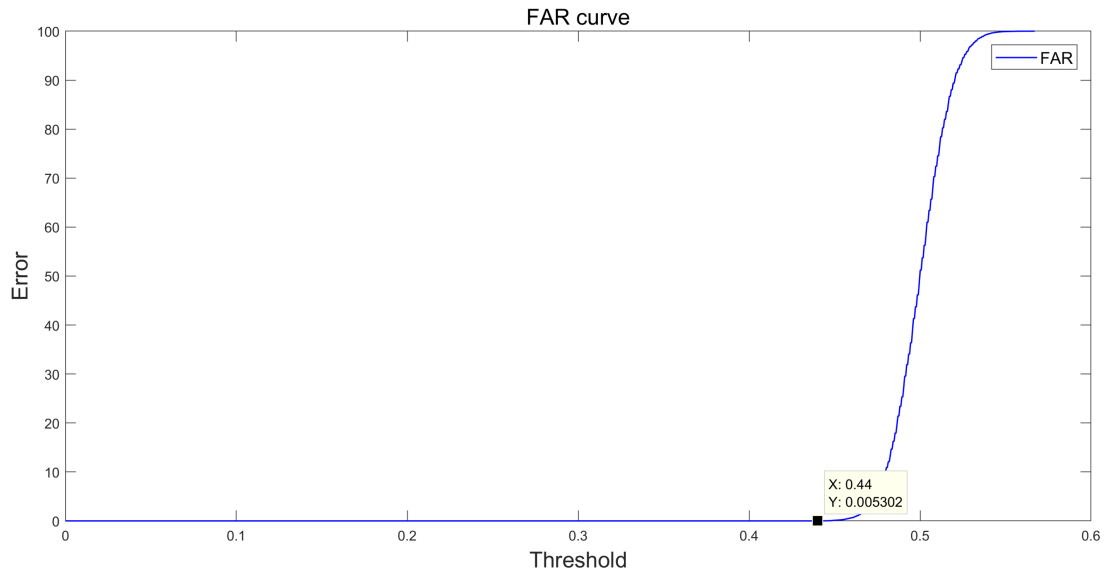
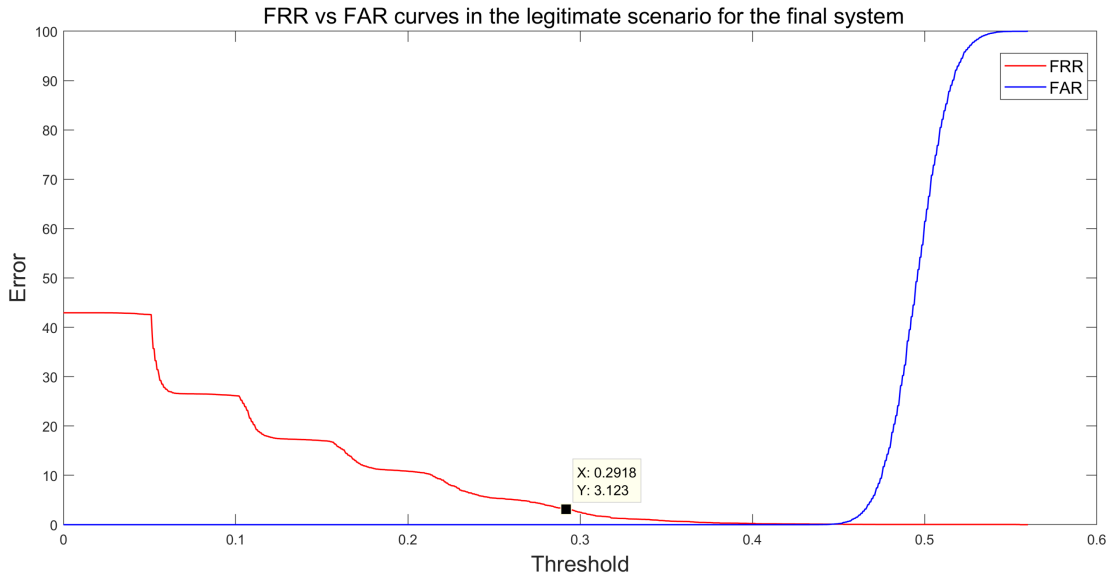
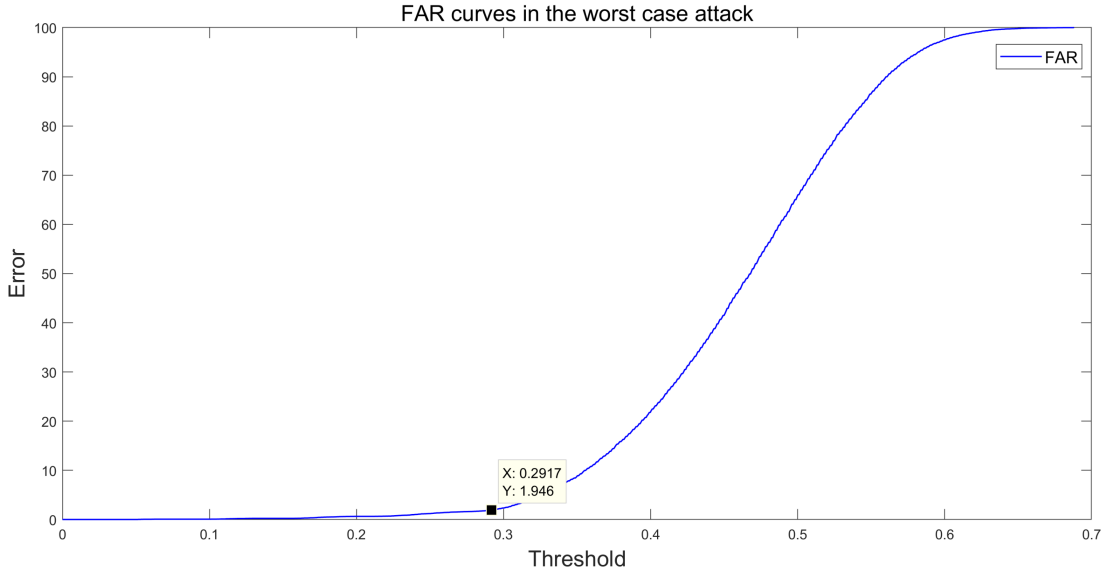


Figure 6.25: FAR curve of the proposed privacy-preserving speaker verification system for the stolen token attack using VoxCeleb1 test set. FAR=0 at the EER threshold.



(a) FRR and FAR curves in the legitimate scenario.



(b) FAR curve in the worst case attack

Figure 6.26: FAR and FRR curves of the privacy-preserving speaker verification system based on corrected cancelable x-vectors in the legitimate and the worst case scenarios using VoxCeleb1 test set.

6.6 Summary of The Results

In this chapter, experimental results were presented for each module in the proposed system: baseline x-vectors (without protection), binary x-vectors, cancelable x-vectors, and cancelable x-vectors with error correction code.

The biometric performance of speaker verification systems based on unprotected (baseline) x-vectors extracted using a TDNN was reported using different back-end scoring. For the baseline systems 1 and 2 where the x-vectors were processed with mean centering, length normalization, and LDA, the EER obtained was 3.12% using PLDA and 5.5% using cosine as back-end scoring. For the baseline system 3 where the cosine distance was used as back-end scoring without applying the LDA and the normalization process, the EER obtained was 8.18%. For our work, we used the x-vectors extracted using system 3 as a baseline to avoid protecting more biometric information such as that contained in the PLDA. However, our goal was to develop a privacy-preserving speaker verification system based on cancelable x-vectors system that maintains the biometric performance of the best baseline x-vectors system 1 (PLDA as back-end).

Therefore, we proposed an approach based on deep neural nets autoencoder trained to transform the x-vector embeddings into binary representations. The autoencoder was composed of an encoder trained to binarize the x-vector embedding and a decoder trained to reconstruct the x-vector from the binary representation. The EER of speaker verification system based on binary x-vectors extracted using the autoencoder was 3.66% compared to the baseline systems with EER = 3.12%,5.5%, and 8.8%. The binarization of speaker embedding using the autoencoder model maintains the performance compared to the baseline systems.

Then, cancelable x-vectors are generated by transforming the binary x-vectors with the shuffling scheme. This transformation allows achieving the privacy requirements and maintaining the biometric performance in the legitimate scenario compared to the baseline systems. For speaker verification based on cancelable x-vectors, the EER obtained was 0.1%. However, degradation was reported in terms of FAR for the stolen key scenario. Compared to the best baseline x-vectors system 3 where the EER=3.12%, using the cancelable x-vectors, at the threshold corresponds to FRR=3.12%, the FAR=4.39% in the stolen key scenario.

To improve the robustness of the system against the stolen key attacks, the idea was to pass the cancelable x-vector through an error-correcting code to manage the biometric variability.
For the proposed system using four enrollment utterances, results show that the speaker verification system based on corrected cancelable x-vectors system achieves the privacy requirements and outperforms the biometric performance of the baseline (without protection) x-vector systems. An EER=0.08% was obtained compared to EER=3.12% for the baseline x-vectors. Also, the system is robust against stolen biometric, stolen token, and brute force attacks with a FAR=0. In addition, due to the combination of shuffling scheme and the error correction code, the proposed system is robust to the stolen shuffling key scenario. For the baseline x-vectors system based on PLDA as back-end scoring, the biometric performance in terms of EER=3.12% (FAR=FRR=3.12%). For the proposed privacy-preserving x-vector system, at FRR=3.12%, the FAR=0 in the legitimate scenario and the FAR=1.94% for the stolen shuffling key scenario. The proposed system outperforms the performance of the baseline system in the legitimate and the stolen key scenarios.

For the system based on one utterance during the enrollment, the system outperforms the baseline system in the legitimate scenario with EER=0.1%. However, a slight degradation in terms of FAR was observed for the stolen key scenario. FAR=4.1% was reported compared to 3.12% for the baseline system.

6.7 Chapter Summary and Conclusions

In this chapter, we have proposed a speaker verification system based on cancelable x-vectors that performs the biometric verification while preserving user privacy. Biometric template protection was performed by first transforming the x-vectors into binary representations using an autoencoder on top of the TDNN. Then, a cancelable x-vector was generated by protecting the binary representation with the shuffling scheme. Next, the Reed-Solomon error-correction code was applied to the cancelable x-vectors to improve the biometric performance and the security of the system.

The proposed system was evaluated according to the requirements of ISO/IEC IS 24745 on biometric information protection; biometric performance, revocability, unlinkability, and irreversibility. Furthermore, the robustness to different attack scenarios has been analyzed. In order to make our research reproducible, the evaluations were performed using public VoxCeleb databases following a clear protocol. The main findings of this chapter can be summarized in the followings points:

- The autoencoder-based binarization approach transforms the x-vector into a binary representation without a loss in biometric performance. The EER of speaker verification system based on the binary x-vectors was 3.66% compared to the baseline x-vectors systems with EER equal to 3.12%, 5.5%, and 8.18%.

- The proposed cancelable x-vectors speaker verification system outperforms the biometric performance of the baseline x-vectors systems. An EER equal or lower to 0.1% is achieved, showing a 97% relative improvement compared to the best baseline x-vectors system (EER=3.12%).

- The proposed system performs speaker verification without revealing the user's biometric information. Only protected x-vectors are stored in the server or handled during the verification phase.

- Revocability is achieved with the use of the shuffling scheme. A new cancelable x-vector can be generated by changing the shuffling key of the user.

- Unlinkability is also achieved. Cancelable x-vectors are unlinkable which avoids cross-matching attacks.
- The proposed system is robust against stolen biometric, stolen token, and brute force attacks with a FAR=0.
- Unlike the systems proposed in chapter 4 and 5, the cancelable x-vectors system based on four utterances for the enrollment is robust to the stolen shuffling key scenario. This was achieved through the binarization approach that maintains the biometric performance and the combination of the cancelable scheme with the error-correcting code.

Chapter 7

Conclusions and Future Research

Contents

7.1 Summary

In this thesis, we addressed the problem of privacy-preserving and security for speaker verification systems. We developed biometric protection schemes for performing speaker verification in a protected domain that preserves the privacy of user's biometric information and improves the robustness of the biometric system. We considered the issue of privacy preservation in the context of three speaker verification systems based on: Gaussian mixture models (GMM), i-vector, and x-vector for speaker modeling. We proposed biometric protection schemes based on the binarization of the speaker representation and its protection using a cancelable scheme to create privacy-preserving biometric systems. Regarding state of the art, most cancelable schemes applied in order to preserve privacy introduce degradation in terms of biometric performance compared to the non-protected system. The proposed privacy-preserving speaker verification systems achieve the privacy and security requirements while maintaining the biometric performance. In addition, to contribute reproducible research and allow comparisons with other approaches,

common and standardized protocols with public and available databases have been used in the experimental evaluations. We summarize the chapters below.

Chapter 1 introduced the privacy issues related to speaker verification systems, the standard ISO/IEC IS 24745 on biometric information protection, the motivation and objectives of the thesis, and the research contributions originated from this thesis.

Chapter 2 summarized the most relevant works related to the research developed in the thesis. We presented the existing approaches for speaker verification systems and the vulnerabilities of these systems in terms of privacy and security. Then, we reviewed the state-of-the-art of researches that address biometric information protection for speaker verification systems.

Chapter 3 described the evaluation of an audio-visual biometric system against presentation attack-based on 3D talking head created from the 2D image and the voice recording of the target user. The evaluation outlined that the fusion of speaker verification system with face modality is not sufficient to achieve secure authentication. With the advancement in the generation of 3D talking head, an attacker can use a 2D image of the target user and perform a 3D facial reconstruction able to bypass the anti-spoofing detectors. This evaluation served as motivation for developing privacy-preserving speaker verification systems.

Chapter 4 presented a biometric protection scheme to develop a privacy-preserving speaker verification system based on Gaussian mixture model. The proposed scheme includes two steps: the representation of acoustic features with a binary representation and then the protection of the binary template with the shuffling scheme. The privacy-preserving system was evaluated according to the requirements of biometric information protection described in ISO/IEC IS 24745 using a text-dependent RSR2015 database. Results show that the proposed system achieves the privacy requirements (revocability, unlinkability, and irreversibility) while maintaining the biometric performance. An improvement in biometric performance in terms of EER compared to the baseline GMM system was reported. As example using the female subset, for target-correct/impostor correct trials of RSR2015 databases, the EER for the baseline system was 1.98% which improved to 0.01% using the protected system. Moreover, the proposed privacy-preserving GMM system is robust against different attack scenarios. For stolen biometric

attack, a false acceptance rate equal to 0 was reported.

Chapter 5 presented a biometric protection scheme for a privacy-preserving speaker verification system based on i-vectors. The scheme includes two steps: i-vector binarization using the thresholding method and the protection of the binary i-vector with the shuffling scheme. The proposed system performs speaker verification without revealing the speaker's voice information to the access server, either during enrollment or during the verification phase. We also demonstrate that this protection scheme could operate to achieve privacy for speaker verification systems based on x-vectors. The proposed systems were evaluated using the RSR2015 text-dependent database and SRE16 text-independent database for the system based on i-vectors and using the VoxCeleb text-independent database for the system based on x-vectors. Compared to the majority of research on voice biometric protection, the proposed privacy-preserving system made it possible to simultaneously achieve privacy requirements and preserve the biometric verification performance. The proposed system improves the biometric performance compared to the unprotected system. Moreover, due to the shuffling scheme, the protected i-vectors are revocable. In case the biometric data or the shuffling key is stolen, different protected i-vectors could be generated from the same voice sample without the possibility to be linked. In addition, the protected system has a good level of security against different attack scenarios. FAR=0 has been reported for brute force attacks, stolen tokens, and stolen biometric attacks.

The main weakness of the privacy-preserving schemes used to protected speaker verification systems based on GMM and i-vectors presented in chapter 4 and 5 respectively is the low resistance to the worst-case attack (stolen shuffling key scenario). In case the shuffling key is stolen, the performance of privacy-preserving systems degrades compared to the baseline systems. This degradation is linked to the loss of biometric performance caused by transforming the speaker biometric reference into a binary representation before applying the shuffling scheme.

In chapter 6, a novel approach for binarizing speaker representation while maintaining the biometric performance was presented. This approach is based on deep neural nets autoencoder trained to transform the x-vector embeddings into binary representations. The autoencoder is composed of an encoder trained to binarize the x-vector embedding and a decoder trained to reconstruct the x-vector

from the binary representation. In contrast to the threshold-based binarization method used in chapter 5 which degrades the biometric performance, binarization of speaker embedding using the autoencoder model maintains the performance obtained with the baseline system. The EER of speaker verification system using the binary representation extracted using the autoencoder was 3.66% compared to the baseline system with EER = 3.12% and 5.5% using PLDA and cosine as back-end scoring respectively. In addition, binarization based on the autoencoder method makes it possible to control the dimension of the binary representation.

This binarization approach was then used to develop a privacy-preserving system based on x-vectors. Protection of x-vector was first performed by transforming it into binary representations using an autoencoder on top of the TDNN. Then, a cancelable x-vector is generated by transforming the binary representation with the shuffling scheme. This transformation allows achieving revocability in case the shuffling key or the cancelable template is compromised. Next, the idea was to pass the cancelable x-vector through a Reed-Solomon error-correction code to manage the intra-variability which allows improving the FAR in the stolen key scenario while maintaining the performance in terms of FRR in the legitimate scenario.

The privacy-preserving speaker verification system based on protected x-vectors system was evaluated using the text-independent VoxCeleb database. The cancelable x-vectors system achieves the privacy requirements and outperforms the biometric performance of the baseline x-vector system. An EER=0.1% was obtained compared to EER=3.12% for the baseline x-vectors. The proposed system is robust against stolen biometric, stolen token, and brute force attacks with a FAR=0. In addition, due to the binarization method and the combination of shuffling scheme and the RS error correction code, the cancelable x-vectors system is robust to the stolen shuffling key scenario. For the baseline x-vectors system based on PLDA as back-end scoring, the biometric performance in terms of EER=3.12% (FAR=FRR=3.12%). For the proposed privacy-preserving x-vector system based on four enrollment utterances, at FRR=3.12%, the FAR=0 in the legitimate scenario and the FAR=1.94% for the stolen shuffling key scenario. The proposed system outperforms the performance of the baseline system in the legitimate and the stolen key scenarios.

7.2 Future Research Directions

Suggested future research work resulting from this thesis can be summarized as follows:

- Exploitation of the speaker model binarization approach based on autoencoder proposed in chapter 6 to develop a privacy-preserving speaker verification system based on homomorphic encryption schemes. Homomorphic encryption is successfully applied to preserve privacy for speaker verification while maintaining biometric performance. However, the computational overhead incurred by processing speech data in the encrypted domain is substantial. To reduce the computational overhead, [Nautsch et al., 2019] propose to operate with binary speaker representation. They demonstrate that using binary representations decreases the computation time required for biometric comparisons in the encrypted domain. Therefore, we believe that the binarization method described in chapter 6 could be used to improve such systems. The autoencoder-based binarization transforms the speaker biometric reference into a binary representation without a loss in biometric performance. Also, it makes possible to control the dimension of binary representation.

- Development of privacy-preserving audio-visual biometric recognition system. Using the autoencoder on top of the TDNN, the voice characteristics could be represented by a binary vector with a dimension chosen according to the dimension of the face representation. This allows to map both modalities into a single representation space and apply biometric protection schemes.

- The proposed cancelable schemes in this thesis address the protection of speaker verification systems based on x-vectors or i-vectors using cosine distance as back-end scoring. Future research could be the development of cancelable scheme dedicated to the protection of speaker verification systems using log-likelihood ratio scores from probabilistic linear discriminant analysis (PLDA).

Bibliography

[Acar et al., 2018] Acar, A., Aksu, H., Uluagac, A. S., and Conti, M. (2018). A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (CSUR)*, 51(4):1–35.

[Alegre et al., 2012] Alegre, F., Vipperla, R., Evans, N., and Fauve, B. (2012). On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals. In *2012 Proceedings of the 20th european signal processing conference (EUSIPCO)*, pages 36–40. IEEE.

[Aliasgari et al., 2017] Aliasgari, M., Blanton, M., and Bayatbabolghani, F. (2017). Secure computation of hidden markov models and secure floating-point arithmetic in the malicious model. *International Journal of Information Security*, 16(6):577–601.

[Anguera and Bonastre, 2010] Anguera, X. and Bonastre, J.-F. (2010). A novel speaker binary key derived from anchor models. In *Eleventh Annual Conference of the International Speech Communication Association*.

[Anguera and Bonastre, 2011] Anguera, X. and Bonastre, J.-F. (2011). Fast speaker diarization based on binary keys. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4428–4431. IEEE.

[Billeb et al., 2015] Billeb, S., Rathgeb, C., Reininger, H., Kasper, K., and Busch, C. (2015). Biometric template protection for speaker recognition based on universal background models. *IET Biometrics*, 4(2):116–126.

[Bonastre et al., 2011] Bonastre, J.-F., Anguera, X., Sierra, G. H., and Bousquet, P.-M. (2011). Speaker modeling using local binary decisions. In *Twelfth Annual Conference of the International Speech Communication Association*.

[Boufounos and Rane, 2011] Boufounos, P. and Rane, S. (2011). Secure binary embeddings for privacy preserving nearest neighbors. In *2011 IEEE international Workshop on information Forensics and security*, pages 1–6. IEEE.

[Cai et al., 2018] Cai, W., Doshi, A., and Valle, R. (2018). Attacking speaker recognition with deep generative models. *arXiv preprint arXiv:1801.02384*.

[Campbell et al., 2006] Campbell, W. M., Sturim, D. E., Reynolds, D. A., and Solomonoff, A. (2006). Svm based speaker verification using a gmm supervector kernel and nap variability compensation. In *2006 IEEE International conference on acoustics speech and signal processing proceedings*, volume 1, pages I–I. IEEE.

[Chee et al., 2018] Chee, K.-Y., Jin, Z., Cai, D., Li, M., Yap, W.-S., Lai, Y.-L., and Goi, B.-M. (2018). Cancellable speech template via random binary orthogonal matrices projection hashing. *Pattern Recognition*, 76:273–287.

[Chung et al., 2018] Chung, J. S., Nagrani, A., and Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.

[Clarke, 2002] Clarke, C. (2002). Reed-solomon error correction, bbc r&d white paper№ 031.

[De Leon et al., 2012] De Leon, P. L., Pucher, M., Yamagishi, J., Hernaez, I., and Saratxaga, I. (2012). Evaluation of speaker verification security and detection of hmm-based synthetic speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8):2280–2290.

[Dehak et al., 2010] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

[Delgado et al., 2015] Delgado, H., Anguera, X., Fredouille, C., and Serrano, J. (2015). Fast single-and cross-show speaker diarization using binary key speaker

modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12):2286–2297.

[Dodis et al., 2004] Dodis, Y., Reyzin, L., and Smith, A. (2004). Fuzzy extractors: How to generate strong keys from biometrics and other noisy data. In *International conference on the theory and applications of cryptographic techniques*, pages 523–540. Springer.

[Ergünay et al., 2015] Ergünay, S. K., Khoury, E., Lazaridis, A., and Marcel, S. (2015). On the vulnerability of speaker verification to realistic voice spoofing. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6. IEEE.

[European Parliament and Council, 2016] European Parliament and Council (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[Evans et al., 2013] Evans, N. W., Kinnunen, T., and Yamagishi, J. (2013). Spoofing and countermeasures for automatic speaker verification. In *Interspeech*, pages 925–929.

[Farrús Cabeceran et al., 2010] Farrús Cabeceran, M., Wagner, M., Erro Eslava, D., and Hernando Pericás, F. J. (2010). Automatic speaker recognition as a measurement of voice imitation and conversion. *The Intenational Journal of Speech. Language and the Law*, 1(17):119–142.

[Garcia et al., 2020] Garcia, D., McCree, A., Snyder, D., and Sell, G. (2020). Jhuhltcoe system for the voxsrc speaker recognition challenge. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7559–7563. IEEE.

[Garcia-Romero and Espy-Wilson, 2011] Garcia-Romero, D. and Espy-Wilson, C. Y. (2011). Analysis of i-vector length normalization in speaker recognition systems. In *Twelfth annual conference of the international speech communication association*.

[Gentry, 2009] Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 169–178.

[Gomez-Barrero et al., 2017] Gomez-Barrero, M., Galbally, J., Rathgeb, C., and Busch, C. (2017). General framework to evaluate unlinkability in biometric template protection systems. *IEEE Transactions on Information Forensics and Security*, 13(6):1406–1420.

[Gómez García et al., 2015] Gómez García, J. A., Moro Velázquez, L., Godino Llorente, J. I., and Castellanos Domínguez, G. (2015). Automatic age detection in normal and pathological voice.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

[Harb and Chen, 2005] Harb, H. and Chen, L. (2005). Voice-based gender identification in multimedia applications. *Journal of intelligent information systems*, 24(2):179–198.

[Hautamäki et al., 2015] Hautamäki, R. G., Kinnunen, T., Hautamäki, V., and Laukkanen, A.-M. (2015). Automatic versus human speaker verification: The case of voice mimicry. *Speech Communication*, 72:13–31.

[Heigold et al., 2016] Heigold, G., Moreno, I., Bengio, S., and Shazeer, N. (2016). End-to-end text-dependent speaker verification. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5115–5119. IEEE.

[Hmani et al.,] Hmani, M. A., Petrovska-Delacrétaz, D., and Dorizzi, B. Locality preserving binary face representations using auto-encoders. *IET Biometrics*, submitted.

[Hunt and Black, 1996] Hunt, A. J. and Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 373–376. IEEE.

[Inthavisas and Lopresti, 2012] Inthavisas, K. and Lopresti, D. (2012). Secure speech biometric templates for user authentication. *IET biometrics*, 1(1):46–54.

[ISO/IEC 30107, 2017] ISO/IEC 30107 (2017). *ISO/IEC 30107-3:2017 Technologies de l'information — Détection d'attaque de présentation en biométrie*. International Organization for Standardization.

[ISO/IEC JTC1 SC27 Security Techniques, 2011] ISO/IEC JTC1 SC27 Security Techniques (2011). *ISO/IEC 24745:2011. Information Technology - Security Techniques - Biometric Information Protection*. International Organization for Standardization.

[Jain et al., 2004] Jain, A. K., Ross, A., and Prabhakar, S. (2004). An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, 14(1):4–20.

[Jeancolas et al., 2019] Jeancolas, L., Mangone, G., Corvol, J.-C., Vidailhet, M., Lehericy, S., Benkelfat, B.-E., Benali, H., and Petrovska-Delacretaz, D. (2019). Comparison of telephone recordings and professional microphone recordings for early detection of parkinson's disease, using mel-frequency cepstral coefficients with gaussian mixture models. In *INTERSPEECH 2019: 20th annual conference of the International Speech Communication Association*, pages 3033–3037. International Speech Communication Association (ISCA).

[Jiménez and Raj, 2017] Jiménez, A. and Raj, B. (2017). A two factor transformation for speaker verification through 1 comparison. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE.

[Juels and Sudan, 2006] Juels, A. and Sudan, M. (2006). A fuzzy vault scheme. *Designs, Codes and Cryptography*, 38(2):237–257.

[Juels and Wattenberg, 1999] Juels, A. and Wattenberg, M. (1999). A fuzzy commitment scheme. In *Proceedings of the 6th ACM conference on Computer and communications security*, pages 28–36.

[Jung et al., 2018] Jung, J.-W., Heo, H.-S., Yang, I.-H., Shim, H.-J., and Yu, H.-J. (2018). Avoiding speaker overfitting in end-to-end dnns using raw waveform for text-independent speaker verification. *extraction*, 8(12):23–24.

[Kanade et al., 2012] Kanade, S. G., Petrovska-Delacrétaz, D., and Dorizzi, B. (2012). Enhancing information security and privacy by combining biometrics with cryptography. *Synthesis Lectures on Information Security, Privacy, and Trust*, 3(1):1–140.

[Kanagasundaram et al., 2011] Kanagasundaram, A., Vogt, R., Dean, D., Sridharan, S., and Mason, M. (2011). I-vector based speaker recognition on short utterances. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, pages 2341–2344. International Speech Communication Association.

[Kaneko et al., 2017] Kaneko, T., Kameoka, H., Hojo, N., Ijima, Y., Hiramatsu, K., and Kashino, K. (2017). Generative adversarial network-based postfilter for statistical parametric speech synthesis. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4910–4914. IEEE.

[Keating et al., 1994] Keating, P. A., Byrd, D., Flemming, E., and Todaka, Y. (1994). Phonetic analyses of word and segment variation using the timit corpus of american english. *Speech Communication*, 14(2):131–142.

[Kenny, 2010] Kenny, P. (2010). Bayesian speaker verification with heavy-tailed priors. In *Odyssey*, volume 14.

[Kenny et al., 2008] Kenny, P., Ouellet, P., Dehak, N., Gupta, V., and Dumouchel, P. (2008). A study of interspeaker variability in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5):980–988.

[Kenny et al., 2014] Kenny, P., Stafylakis, T., Ouellet, P., Gupta, V., and Alam, M. J. (2014). Deep neural networks for extracting baum-welch statistics for speaker recognition. In *Odyssey*, volume 2014, pages 293–298.

[Kinnunen et al., 2017] Kinnunen, T., Sahidullah, M., Delgado, H., Todisco, M., Evans, N., Yamagishi, J., and Lee, K. A. (2017). The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection.

[Kinnunen et al., 2012] Kinnunen, T., Wu, Z.-Z., Lee, K. A., Sedlak, F., Chng, E. S., and Li, H. (2012). Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4401–4404. IEEE.

[Kumar et al., 2020] Kumar, N. et al. (2020). Cancelable biometrics: A comprehensive survey. *Artificial Intelligence Review*, 53(5):3403–3446.

[Larcher et al., 2014] Larcher, A., Lee, K. A., Ma, B., and Li, H. (2014). Text-dependent speaker verification: Classifiers, databases and rsr2015. *Speech Communication*, 60:56–77.

[Lau et al., 2005] Lau, Y. W., Tran, D., and Wagner, M. (2005). Testing voice mimicry with the yoho speaker verification corpus. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 15–21. Springer.

[Lawson et al., 2011] Lawson, A., Vabishchevich, P., Huggins, M., Ardis, P., Battles, B., and Stauffer, A. (2011). Survey and evaluation of acoustic features for speaker recognition. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5444–5447. IEEE.

[Lei et al., 2014] Lei, Y., Scheffer, N., Ferrer, L., and McLaren, M. (2014). A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1695–1699. IEEE.

[Li et al., 2016] Li, L., Xing, C., Wang, D., Yu, K., and Zheng, T. F. (2016). Binary speaker embedding. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–4. IEEE.

[MacWilliams and Sloane, 1977] MacWilliams, F. J. and Sloane, N. J. A. (1977). *The theory of error correcting codes*, volume 16. Elsevier.

[Mandalapu et al., 2021] Mandalapu, H., Busch, C., and Ramachandra, R. (2021). Multilingual voice impersonation dataset and evaluation. In *Proceeding of the 3rd International Conference on Intelligent Technologies and Applications (INTAP)*. Springer.

[Matrouf et al., 2006] Matrouf, D., Bonastre, J.-F., and Fredouille, C. (2006). Effect of speech transformation on impostor acceptance. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE.

[McLaren et al., 2015] McLaren, M., Lei, Y., and Ferrer, L. (2015). Advances in deep neural network approaches to speaker recognition. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4814–4818. IEEE.

[Mehri et al., 2016] Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., Courville, A., and Bengio, Y. (2016). Samplernn: An unconditional end-to-end neural audio generation model. *arXiv preprint arXiv:1612.07837*.

[Nagrani et al., 2017] Nagrani, A., Chung, J. S., and Zisserman, A. (2017). Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.

[Nautsch et al., 2018] Nautsch, A., Isadskiy, S., Kolberg, J., Gomez-Barrero, M., and Busch, C. (2018). Homomorphic encryption for speaker recognition: Protection of biometric templates and vendor model parameters. *arXiv preprint arXiv:1803.03559*.

[Nautsch et al., 2019] Nautsch, A., Patino, J., Treiber, A., Stafylakis, T., Mizera, P., Todisco, M., Schneider, T., and Evans, N. (2019). Privacy-preserving speaker recognition with cohort score normalisation. *arXiv preprint arXiv:1907.03454*.

[Paillier, 1999] Paillier, P. (1999). Public-key cryptosystems based on composite degree residuosity classes. In *International conference on the theory and applications of cryptographic techniques*, pages 223–238. Springer.

[Palaz et al., 2015] Palaz, D., Collobert, R., et al. (2015). Analysis of cnn-based speech recognition system using raw speech as input. Technical report, Idiap.

[Panjwani and Prakash, 2014] Panjwani, S. and Prakash, A. (2014). Crowdsourcing attacks on biometric systems. In *10th Symposium On Usable Privacy and Security ({SOUPS} 2014)*, pages 257–269.

[Pathak and Raj, 2011] Pathak, M. A. and Raj, B. (2011). Privacy preserving speaker verification using adapted gmms. In *Twelfth Annual Conference of the International Speech Communication Association*.

[Pathak and Raj, 2012a] Pathak, M. A. and Raj, B. (2012a). Privacy-preserving speaker verification and identification using gaussian mixture models. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(2):397–406.

[Pathak and Raj, 2012b] Pathak, M. A. and Raj, B. (2012b). Privacy-preserving speaker verification as password matching. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1849–1852. IEEE.

[Paulini et al., 2016] Paulini, M., Rathgeb, C., Nautsch, A., Reichau, H., Reininger, H., and Busch, C. (2016). Multi-bit allocation: Preparing voice biometrics for template protection. In *Odyssey*, pages 291–296.

[Peddinti et al., 2015] Peddinti, V., Povey, D., and Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth annual conference of the international speech communication association*.

[Portêlo et al., 2014a] Portêlo, J., Raj, B., Abad, A., and Trancoso, I. (2014a). Privacy-preserving speaker verification using garbled gmms. In *2014 22nd European Signal Processing Conference (EUSIPCO)*, pages 2070–2074. IEEE.

[Portêlo et al., 2014b] Portêlo, J., Raj, B., Abad, A., and Trancoso, I. (2014b). Privacy-preserving speaker verification using secure binary embeddings. In *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1268–1272. IEEE.

[Povey et al., 2011] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society.

[Raj et al., 2019] Raj, D., Snyder, D., Povey, D., and Khudanpur, S. (2019). Probing the information encoded in x-vectors. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 726–733. IEEE.

[Rathgeb and Uhl, 2011] Rathgeb, C. and Uhl, A. (2011). A survey on biometric cryptosystems and cancelable biometrics. *EURASIP Journal on Information Security*, 2011(1):1–25.

[Reynolds et al., 2000] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41.

[Rosenberger, 2018] Rosenberger, C. (2018). Evaluation of biometric template protection schemes based on a transformation. In *ICISSP*, pages 216–224.

[Sadjadi et al., 2017] Sadjadi, S. O., Kheyrkhah, T., Tong, A., Greenberg, C. S., Reynolds, D. A., Singer, E., Mason, L. P., Hernandez-Cordero, J., et al. (2017). The 2016 nist speaker recognition evaluation. In *Interspeech*, pages 1353–1357.

[Sadjadi et al., 2013] Sadjadi, S. O., Slaney, M., and Heck, L. (2013). Msr identity toolbox v1. 0: A matlab toolbox for speaker-recognition research. *Speech and Language Processing Technical Committee Newsletter*, 1(4).

[Sahidullah et al., 2019] Sahidullah, M., Delgado, H., Todisco, M., Kinnunen, T., Evans, N., Yamagishi, J., and Lee, K.-A. (2019). Introduction to voice presen-

tation attack detection and recent advances. In *Handbook of Biometric Anti-Spoofing*, pages 321–361. Springer.

[Senior and Fructuoso, 2016] Senior, A. W. and Fructuoso, J. G. (2016). Deep networks for unit selection speech synthesis. US Patent 9,460,704.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

[Shen et al., 2018] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE.

[Snyder et al., 2015] Snyder, D., Chen, G., and Povey, D. (2015). Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

[Snyder et al., 2017] Snyder, D., Garcia-Romero, D., Povey, D., and Khudanpur, S. (2017). Deep neural network embeddings for text-independent speaker verification. In *Interspeech*, pages 999–1003.

[Snyder et al., 2018] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE.

[Snyder et al., 2016] Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y., and Khudanpur, S. (2016). Deep neural network-based speaker embeddings for end-to-end speaker verification. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 165–170. IEEE.

[Taylor et al., 2017] Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A. G., Hodgins, J., and Matthews, I. (2017). A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(4):1–11.

[Teoh and Chong, 2010] Teoh, A. B. J. and Chong, L.-Y. (2010). Secure speech template protection in speaker verification system. *Speech communication*, 52(2):150–163.

[Teoh and Yuang, 2007] Teoh, A. B. J. and Yuang, C. T. (2007). Cancelable biometrics realization with multispace random projections. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(5):1096–1106.

[Tissier et al., 2019] Tissier, J., Gravier, C., and Habrard, A. (2019). Near-lossless binarization of word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7104–7111.

[Todisco et al., 2019] Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T., and Lee, K. A. (2019). Asvspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*.

[Treiber et al., 2019] Treiber, A., Nautsch, A., Kolberg, J., Schneider, T., and Busch, C. (2019). Privacy-preserving plda speaker verification using outsourced secure computation. *Speech Communication*, 114:60–71.

[Tremlet, 2016] Tremlet, C. (2016). Embedded secure element for authentication, storage and transaction within a mobile terminal. US Patent 9,436,940.

[Variani et al., 2014] Variani, E., Lei, X., McDermott, E., Moreno, I. L., and Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4052–4056. IEEE.

[Villalba et al., 2020] Villalba, J., Garcia-Romero, D., Chen, N., Sell, G., Borgstrom, J., McCree, A., Garcia-Perera, L., Kataria, S., Nidadavolu, P. S., Torres-Carrasquillo, P. A., et al. (2020). Advances in speaker recognition for telephone and audio-visual data: the jhu-mit submission for nist sre19. In *Proceedings of Odyssey*.

[Villalba and Lleida, 2010] Villalba, J. and Lleida, E. (2010). Speaker verification performance degradation against spoofing and tampering attacks. In *FALA workshop*, pages 131–134.

[Villalba and Lleida, 2011] Villalba, J. and Lleida, E. (2011). Detecting replay attacks from far-field recordings on speaker verification systems. In *European Workshop on Biometrics and Identity Management*, pages 274–285. Springer.

[Wu et al., 2015] Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilçi, C., Sahidullah, M., and Sizov, A. (2015). Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Sixteenth annual conference of the international speech communication association.*

[Wu and Li, 2013] Wu, Z. and Li, H. (2013). Voice conversion and spoofing attack on speaker verification systems. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–9. IEEE.

[Xu et al., 2016] Xu, Y., Price, T., Frahm, J.-M., and Monrose, F. (2016). Virtual u: Defeating face liveness detection by building virtual models from your public photos. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pages 497–512.

[Yamagishi et al., 2021] Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., Liu, X., Lee, K. A., Kinnunen, T., Evans, N., et al. (2021). Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537.*

[Yao, 1982] Yao, A. C. (1982). Protocols for secure computations. In *23rd annual symposium on foundations of computer science (SFCS 1982)*, pages 160–164. IEEE.

[Yoshimura et al., 1999] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. In *Sixth European Conference on Speech Communication and Technology.*

[Zeinali et al., 2019] Zeinali, H., Wang, S., Silnova, A., Matějka, P., and Plchot, O. (2019). But system description to voxceleb speaker recognition challenge 2019. *arXiv preprint arXiv:1910.12592*.

[Zen et al., 2009] Zen, H., Tokuda, K., and Black, A. W. (2009). Statistical parametric speech synthesis. *Speech communication*, 51(11):1039–1064.

[Zhou et al., 2019] Zhou, T., Zhao, Y., Li, J., Gong, Y., and Wu, J. (2019). Cnn with phonetic attention for text-independent speaker verification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 718–725. IEEE.

[Zhu et al., 2012] Zhu, H.-H., He, Q.-H., and Li, Y.-X. (2012). A two-step hybrid approach for voiceprint-biometric template protection. In *2012 International Conference on Machine Learning and Cybernetics*, volume 2, pages 560–565. IEEE.

DOCTORAL SCHOOL

Titre : VERS DES SYSTÈMES DE VÉRIFICATION DE LOCUTEUR ROBUSTES ET PRÉSERVANT LA VIE PRIVÉE

Mots clés : Système de Vérification de Locuteur, Protection de la Vie Privée, Sécurité, Protection des Informations Biométriques.

Résumé : Les systèmes de vérification du locuteur sont une technologie clé dans de nombreux appareils et services tels que les smartphones, les assistants numériques intelligents et les applications bancaires. Pendant la pandémie de COVID-19, les systèmes de contrôle d'accès basés sur des lecteurs d'empreintes digitales ou des claviers augmentent le risque de propagation du virus. Par conséquent, les entreprises repensent maintenant leurs systèmes de contrôle d'accès des employés et envisagent des technologies d'autorisation sans contact, telles que les systèmes de vérification des locuteurs.
Cependant, les systèmes de vérification des locuteurs exigent que le système d'accès stocke les modèles des locuteurs et ait accès aux enregistrements ou aux caractéristiques dérivées des voix des locuteurs lors de l'authentification. Ce processus soulève certaines préoccupations concernant le respect de la vie privée de l'utilisateur et la protection de ces données biométriques sensibles. Un adversaire peut voler les informations biométriques des locuteurs pour usurper l'identité de l'utilisateur authentique et obtenir un accès non autorisé. De plus, lorsqu'il s'agit de données vocales, nous sommes confrontés à des problèmes supplémentaires de confidentialité et de respect de vie privée parce que à partir des données vocales plusieurs informations personnelles liées à l'identité, au sexe, à l'âge ou à l'état de santé du locuteur peuvent être extraites.
Dans ce contexte, la présente thèse de doctorat aborde les problèmes de protection des données biométriques, le respect de vie privée et la sécurité pour les systèmes de vérification du locuteur basés sur les modèles de mélange gaussien, i-vecteur et x-vecteur comme modélisation du locuteur. L'objectif est le développement de systèmes de vérification du locuteur qui effectuent une vérification biométrique tout en respectant la vie privée et la protection des données biométriques de l'utilisateur. Pour cela, nous avons proposé des schémas de protection biométrique afin de répondre aux exigences de protection des données biométriques (révocabilité, diversité, et irréversibilité) décrites dans la norme ISO/IEC IS 24745 et pour améliorer la robustesse des systèmes contre différentes scénarios d'attaques.

Title : TOWARDS ROBUST AND PRIVACY-PRESERVING SPEAKER VERIFICATION SYSTEMS

Keywords : Speaker Verification, Privacy, Security, Biometric Information Protection

Abstract : Speaker verification systems are a key technology in many devices and services like smartphones, intelligent digital assistants, healthcare, and banking applications. Additionally, with the COVID pandemic, access control systems based on fingerprint scanners or keypads increase the risk of virus propagation. Therefore, companies are now rethinking their employee access control systems and considering touchless authorization technologies, such as speaker verification systems. However, speaker verification system requires users to transmit their recordings, features, or models derived from their voice samples without any obfuscation over untrusted public networks which stored and processed them on a cloud-based infrastructure. If the system is compromised, an adversary can use this biometric information to impersonate the genuine user and extract personal information. The voice samples may contain information about the user's gender, accent, ethnicity, and health status which raises several privacy issues. In this context, the present PhD Thesis address the privacy and security issues for speaker verification systems based on Gaussian mixture models (GMM), i-vector, and x-vector as speaker modeling. The objective is the development of speaker verification systems that perform biometric verification while preserving the privacy and the security of the user. To that end, we proposed biometric protection schemes for speaker verification systems to achieve the privacy requirements (revocability, unlinkability, irreversibility) described in the standard ISO/IEC IS 24745 on biometric information protection and to improve the robustness of the systems against different attack scenarios.

Institut Polytechnique de Paris
91120 Palaiseau, France