Advances in Privacy Preservation and Fairness Assessment for Voice Biometrics

Oubaïda Chouchane

Advances in Privacy Preservation and Fairness Assessment for Voice Biometrics

Dissertation

submitted to

Sorbonne Université

in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

Author:
Oubaïda Chouchane

Defended on the
10^{th} *of June 2024*
before a committee composed of:

President	**Prof. Nicholas Evans**, EURECOM, France
Reviewers	**Prof. Patrizio Campisi**, Roma Tre University, Italy
	Prof. Gian Luca Marcialis, University of Cagliari, Italy
Examiners	**Prof. Marta Gomez Barrero**, Universität der Bundeswehr München, Germany
	Prof. Christoph Busch, Hochschule Darmstadt, Germany
	Prof. Nicholas Evans, EURECOM, France
	Prof. Chiara Galdi, EURECOM, France
Co-supervisor	**Prof. Massimiliano Todisco**, EURECOM, France
Thesis advisor	**Prof. Melek Önen**, EURECOM, France

Acknowledgements

I would like to extend my acknowledgment to Prof. Melek Önen for her role as a supervisor and for the academic oversight she provided. My appreciation also goes to my co-supervisor, Prof. Massimiliano Todisco, for his invaluable insights and active engagement that have shaped my research.

A warm thank you to all my Eurecom colleagues, with special mentions to Wanying Ge for the light-hearted jokes that brought much-needed relief, and to Michele Panariello for the dark humor jokes that allowed us to find laughter in our shared struggles. Your camaraderie has made this journey all the more memorable.

Turning to the cornerstone of my life—my family—to whom I owe an immeasurable debt of gratitude. My father, my guiding light, you taught me the principles of "le savoir-vivre", self-reliance, and the courage to uphold my beliefs. Your teachings have molded me into the person I am today, and for this, I am eternally grateful.

My mother, you have been a wellspring of endless love and support. Your sacrifices have been the unspoken forces behind every step. Mom, I know how much this means to you, and I dedicate this achievement to you, a small gift in return for the care you have given me.

My brother, the beacon of passion and humor, you have been a continual source of joy. I am immensely grateful to have you in my life; your spirit, kindness, and the laughter we share leave a lasting imprint on my heart. The moments spent with you are not merely amusing, they are treasures I hold dear.

My sister, your inner and outer beauty, brilliance, and tenderness have enriched my life immeasurably. Your presence is a cherished treasure, and every moment

with you is precious. I am proud to have you in our family.

My second parents, your warmth and prayers have doubled the blessings in my life. It fills my heart with immense happiness to be called your daughter. I feel incredibly fortunate to receive such boundless love from each one of you.

To my partner, best friend, and husband: my eternal love and gratitude know no bounds. Your unwavering support, love, wisdom, understanding, and guidance have been my north star, making this journey possible. You are the reason of my internal peace, and thanks to you, I am a fervent believer in the existence of true love. Merely a few lines here cannot describe the depth of my love and appreciation, for you and for all the cherished members of our family.

The creation of this work has been a transformative journey, and my heartfelt gratitude goes to everyone who has supported, guided, and enriched this experience. It symbolizes not just my academic aspirations but also the impact of the shared knowledge and diverse viewpoints.

I must also recognize my resilience and bravery. Each step was laden with unprecedented challenges and emotional struggles, but my unwavering determination to make meaningful contributions and my refusal to surrender have helped me accomplish this journey.

May this work serve as a symbol of collective insight and diligence, providing a modest contribution to the scientific field.

Antibes, March 2024

Oubaïda Chouchane

Abstract

In recent years, the adoption of Automatic Speaker Verification (ASV) systems has significantly expanded to meet the rising demand for secure and effective identity verification methods. However, this widespread adoption of ASV systems has raised concerns regarding privacy and fairness, necessitating compliance with regulations like the European General Data Protection Regulation (GDPR). The GDPR mandates stringent privacy protection measures for biometric data, including voice, as it is classified as sensitive personal data. Soft biometrics embedded in voice data, such as gender, accent, and emotion, present privacy risks if not adequately safeguarded. Additionally, biases within ASV systems can lead to discriminatory outcomes, violating GDPR principles of fairness and non-discrimination.

This thesis addresses the imperative need to enhance compliance with GDPR principles concerning *data privacy* and *fairness* in voice biometrics applications, extending beyond ASV systems to encompass spoofing countermeasures (CMs) systems. These CMs are integral for fortifying ASV systems against spoofing attacks, ensuring the integrity of the authentication process.

To address these challenges, this thesis makes several contributions. The first contribution is PRIVASP, the privacy-preserving scheme for CMs systems using secure multi-party computation (MPC). PRIVASP ensures the privacy of individuals when using CMs systems and safeguards the intellectual property (IP) of these systems by keeping model parameters private. This solution successfully balances privacy requirements with spoofing detection performance, meeting GDPR *privacy by design* principles.

In adherence to GDPR principles concerning data privacy, the second contribution introduces an innovative approach to safeguarding gender information within ASV systems. This method combines differential privacy (DP) mechanisms with adversarial auto-encoder (AAE) techniques to conceal gender-related information

within speaker embeddings while preserving their utility for speaker verification purposes. Our approach enables the selection of the desired balance between privacy protection and utility by adjusting the privacy budget of the DP mechanism, even after the training process is completed.

Furthermore, in compliance with GDPR principles aimed at ensuring data privacy, non-discrimination and fairness in automated decision-making systems, we developed an approach that carefully balances trade-offs between speaker verification accuracy, gender data privacy, and fairness. This is achieved by fine-tuning the pre-trained model of wav2vec 2.0. While previous work has addressed these challenges individually, our research is the first to concurrently address these three aspects in the context of automatic speaker verification. We also introduce a novel fairness metric tailored for voice biometrics, adapted from facial biometrics, to assess the impact of gender information on ASV system fairness. This innovative metric underscores the significance of fairness evaluation in ASV system development, advocating for *fairness by design* to ensure equitable outcomes across demographic groups.

Finally, in this thesis, we address the absence of international standards for fairness evaluation in biometric systems by assessing three fairness metrics in the context of speaker verification. Our comparative analysis reveals the Gini aggregation rate for biometric equitability (GARBE) metric as the most suitable for evaluating ASV system fairness. We then utilize the GARBE metric to evaluate the fairness of five state-of-the-art ASV systems. This illustrates the need to integrate fairness evaluation into ASV system development processes to achieve equitable and reliable speaker verification systems compliant with regulatory standards.

Contents

List of Figures

List of Tables

List of Abbreviations

ASV	Automatic Speaker Verification
CM	Countermeasure
MPC	Secure Multi-Party Computation
IP	Intellectual Property
DP	Differential Privacy
AAE	Adversarial Auto-Encoder
GARBE	Gini Aggregation Rate for Biometric Equitability
DNN	Deep Neural Network
LPC	Linear Prediction Coding
MFCC	Mel-Frequency Cepstral Coefficients
LFCC	Linear Frequency Cepstral Coefficients
GMM	Gaussian Mixture Model
ResNet	Residual Network
TDNN	time delay neural network
PLDA	Probabilistic Linear Discriminant Analysis
GDPR	General Data Protection Regulation
AI	Artificial Intelligence
EU	European Union
PETs	Privacy-Enhancing Technologies
VC	Voice Conversion
TTS	Text-To-Speech
FAD	fairness activation discrepancy
FR	Face Recognition
FR	Face Recognition
ML	Machine Learning
FETs	fairness enhancing technologies
2PC	Secure Two-Party Computation

GC	Garbled Circuits
UBM	Universal Background Model
HMM	Hidden Markov Model
HE	Homomorphic Encryption
PHE	Partially Homomorphic Encryption
SHE	Somewhat Homomorphic Encryption
FHE	Fully Homomorphic Encryption
BGN	Boneh-Goh-Nissim
BGV	Brakerski-Gentry-Vaikuntanathan
BFV	Brakerski/Fan-Vercauteren
LWE	Learning With Errors
RLWE	Ring Learning With Errors
CKKS	Cheon-Kim-Kim-Song
GDP	Global Differential Privacy
LDP	Local Differential Privacy
PCA	Principle Component Analysis
BN	Bottleneck
ASR	Automatic Speech Recognition
DRL	Disentangled Representations Learning
AE	Auto-Encoder
VAE	Variational Auto-Encode
GAN	Generative Adversarial Networks
VPC	VoicePrivacy Challenge
SAN	Semi-Adversarial Network
CAE	Convolutional Auto-Encoder
IVE	Incremental Variable Elimination
FM	False Match
FNM	False Non Match
EER	Equal Error Rate
FAR	False Acceptance Rate
FRR	False Rejection Rate
DET	Detection Cost Function
minDET	minimal Detection Cost Function
FDR	Fairness Discrepancy Rate

PA	Physical access
LA	Logical access
PAD	Presentation Attack Detection
FAs	False Accepts
FRs	False Rejects
t-DCF	Tandem Detection Cost Function
CQCCs	Constant Q Cepstral Coefficients
DFT	Discrete Fourier Transform
EM	Expectation-Maximization
ReLU	Rectified Linear Unit
Adam	Adaptive moment estimation
MSB	Most Significant Bit
AUC	Area Under the ROC Curve
AAM	Additive Angular Margin
GRL	Gradient Reversal Layer
FPD	False Positive Differential
FND	False Negative Differential
auFDR	Area Under Fairness Discrepancy Rate
IA	Informed attack
uIA	Uninformed attack
CNN	Convolutional Neural Network
IR	Inequity Rate
FFMC	Functional Fairness Measure Criteria
SAP	Self-Attentive Pooling
ASP	Attentive Statistics Pooling
D-TDNN	Densely connected Time Delay Neural Network
CAM	Context-Aware Masking

Publications

1. **Oubaïda Chouchane**, Baptiste Brossier, Jorge Esteban Gamboa Gamboa, Thomas Lardy, Hemlata Tak, Orhan Ermis, Madhu Kamble, Jose Patino, Nicholas Evans, Melek Önen, Massimiliano Todisco, "**Privacy-preserving voice anti-spoofing using secure multi-party computation**," in *Proc. INTERSPEECH 2021*, Brno, Czech Republic, September 2021.

2. **Oubaïda Chouchane**, Michele Panariello, Oualid Zari, Ismet Kerenciler, Imen Chihaoui, Massimiliano Todisco, Melek Önen, "**Differentially Private Adversarial Auto-Encoder to Protect Gender in Voice Biometrics**," in *Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security 2023*, Chicago, United States, June 2023.

3. **Oubaïda Chouchane**, Michele Panariello, Chiara Galdi, Massimiliano Todisco, Nicholas Evans, "**Fairness and Privacy in Voice Biometrics:A Study of Gender Influences Using wav2vec 2.0**," in *BIOSIG 2023*, Darmstadt, Germany, September 2023.

4. **Oubaïda Chouchane**, Christoph Busch, Chiara Galdi, Nicholas Evans, and Massimiliano Todisco, "**A Comparison of Differential Performance Metrics for the Evaluation of Automatic Speaker Verification Fairness**," in *Odyssey 2024*, Québec, Canada, June 2024.

Other work

1. Nan Cheng, Melek Önen, Aikaterini Mitrokotsa, **Oubaïda Chouchane**, and Massimiliano Todisco, Alberto Ibarrondo, "**Nomadic: Normalising Maliciously-Secure Distance with Cosine Similarity for Two-Party Biometric Authentication**," in *ACM ASIACCS 2024*.

In the above manuscript, I proposed the main scenario for computing cosine similarity in a privacy-preserving manner against a malicious server in a multi-party computation setting. My involvement also included discussion, experimentation, and writing the paper.

Chapter 1

Introduction

1.1 Voice Biometrics

In recent years, biometric technology has revolutionized access control. Instead of relying on easily forgotten passwords and managing multiple tokens, authentication of individuals now depends on the use of our physiological and behavioral traits such as fingerprints, faces, irises, voice, signature dynamics, and gait. These traits serve as reliable markers of our identity and are impossible to forget or misplace. The availability of affordable devices like microphones, smartwatches, and cameras has made biometric authentication more accessible in our daily lives, providing a convenient and effective way to secure our data. Among these biometric modalities, voice biometrics, also known as Automatic Speaker Verification (ASV), has emerged as a prominent tool.

The advancements in deep neural networks (DNNs) enhanced accuracy and reliability of ASV systems [1–3], thereby contributing to the widespread adoption of these systems. Several banks [4] and call centers have adopted speaker verification solutions as an alternative to traditional passwords to fortify security measures and provide a more efficient and user-friendly experiences [1] [2].

An ASV system verifies speakers by comparing the new speech data with a reference previously saved in the system's database during enrolment. If no adequate match is found that surpasses a predetermined threshold, the system denies the identity claim.

The ASV system operates in two distinct phases: enrolment phase and verification

[1] https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4OcMZ

[2] www.thebanker.com/Voice-recognition-what-your-bank-needs-1462176012

phase, illustrated in Figure 1.1.

In the enrolment phase, the speech signal is fed into a feature extraction mod-

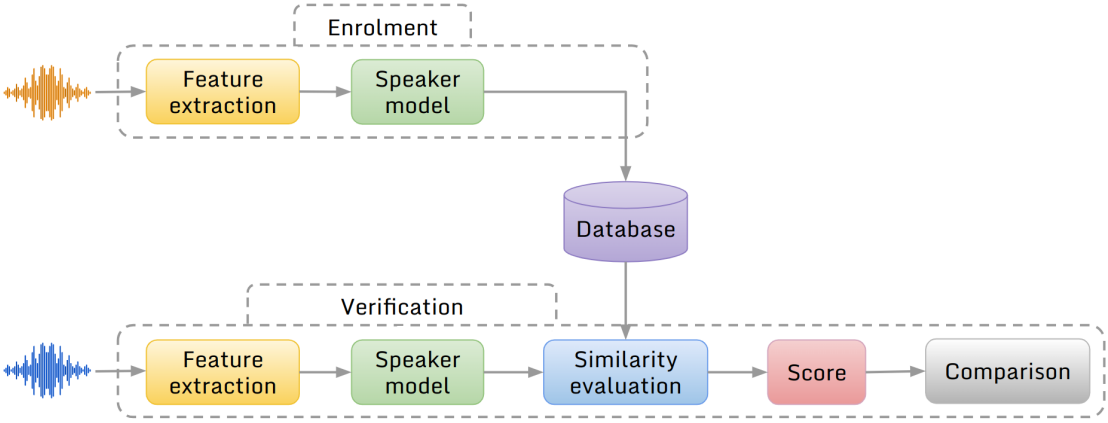


Figure 1.1: Phases of an automatic speaker verification system.

ule that starts by segmenting the speech signal into frames. A frame-level feature extraction technique [5] such as linear prediction coding (LPC), mel-frequency cepstral coefficients (MFCC), linear frequency cepstral coefficients (LFCC), and Mel spectrogram, is then used to capture speaker-specific attributes. For each frame, a feature vector is generated. These features are then aggregated into utterances that encapsulate the temporal dynamics of the speech signal. The utterance-level features are then fed to a speaker modeling module like Gaussian mixture model(GMM) [6], deep neural network (DNN) [7], residual network (ResNet) [8,9], and time delay neural network (TDNN) [1, 10], to learn speaker-specific patterns over time. The speaker modeling procedure is also referred to as the *front-end* stage, where an utterance in time domain or time-frequency domain is mapped to high-dimensional feature vector (i.e. embedding) that represents the identity of the speaker [5]. Finally, the generated speaker model is saved into a database.

During verification, the incoming speech signal undergoes the same feature extraction and modeling process. The new speaker model is then compared to the saved model of the claimed identityusing a similarity measure or dissimilarity measure, such as probabilistic linear discriminant analysis (PLDA), Euclidean distance, and cosine distance [1, 11]. This comparison computation is also known as the *back-end* stage [5]. If there is sufficient correlation, a high similarity score is obtained and the system authenticates the user. Otherwise, the verification of

the speaker is rejected.

In order to enhance their robustness, performance, and computational efficiency, state-of-the-art ASV systems make use of a set of background speakers [12, 13] or cohort speakers [14] as negative examples (i.e. imposter speakers). Models like universal background Model (UGMMs) used a set of background speakers to help the ASV system generalize better to unseen speakers [15, 16]. Recent DNN-based ASV systems use scores of cohort speakers who are close to the genuine speaker, at the back-end for similarity score normalization [5, 15, 17].

Another technique to enhance performance and robustness across different environments or datasets of ASV systems is domain adaptation [18]. ASV systems may encounter domain mismatch problems when dealing with data from "in-the-wild" scenarios, where variability in recording conditions, languages, or speaker demographics can occur. Domain adaptation techniques in ASV involve fine-tuning the parameters of the system, such as feature extraction methods or scoring algorithms, to better align with the characteristics of the target domain.

The ASV systems can be categorized into two main classes: text-dependent and text-independent. In a text-dependent system, it is imperative that the text spoken during the verification matches the text from the enrolment phase. On the other hand, the text-independent system imposes no such requirement, allowing for verification without any predetermined constraints on the phrases uttered by the speaker. Text-independent speaker verification offers greater flexibility compared to the text-dependent approach, as it allows the speaker to spontaneously interact with the system. Despite the additional challenge of training text-independent ASV models, which require longer utterances to achieve higher accuracy, they prove to be more convenient, particularly in scenarios involving spontaneous speech. In this thesis, we focus on text-independent ASV systems.

1.2 Regulatory Requirements

To govern the security and privacy of biometric data, regulations such as the European General Data Protection Regulation (GDPR) [19] and the European artificial intelligence (AI) act [3] [20], have been established. Since May 2018, the process-

[3] https://iapp.org/news/a/biometrics-under-the-eu-ai-act/

ing of biometric data of European union (EU) citizens and residents has been regulated by the GDPR [19]. Being considered the strongest in the world [4], this regulation governs the *processing of personal data* of individuals in the EU. Article 4.1 of the GDPR defines *personal data* as any information directly or indirectly related to identity of individuals. Article 4.2. explains that the term *processing* includes any operation performed on *personal data*, including but not limited to collection, recording, storage, manipulation, transfer, and erasure. Additionally, article 4.14 explicitly categorizes *biometric data* as *personal data*. Besides, biometric data, including voice, is identified as *sensitive personal data* or *special categories of personal data* in Article 9 of the GDPR. The delicacy of processing such data is stressed in Article 5, which mandates the inclusion of appropriate security measures for personal data. Recital 51 as well underscores the potential risk of sensitive personal data processing to fundamental rights and freedoms.

Moreover, the GDPR emphasizes principles of *data protection by design*, detailed in Article 25, which instructs the consideration of the associated risks to the rights and freedoms of individuals, as well as integrating privacy-preservation techniques at an early stage of the processing. Therefore, voice biometric applications such as automatic speaker verification must implement privacy-enhancing technologies (PETs) when storing voice biometrics data in databases or using it during identity verification.

Furthermore, the GDPR requires compliance to *fairness* principle especially by automated decision-making systems, including ASVs. Article 5.1 states out that processing personal data should be fair. Recital 71 further explains that appropriate measures must be included to prevent discrimination and ensure fair outcomes of AI systems. The GDPR aims to protect the right of individuals not only to privacy but also to non-discrimination. Hence, it is imperative to examine the fairness of outcomes produced by ASV systems.

1.3 Privacy and Fairness Issues

In the context of biometrics, privacy is related to the right of individuals to control and protect their biometric data from unauthorized access or misuse. Only authorized entities, that the user has consent to, have the right to process their

[4] `https://www.consilium.europa.eu/en/policies/data-protection/data-protection-regulation/`

biometric data and for biometric recognition tasks only [5]. However, an ASV system might intentionally or unintentionally allows other recognition tasks. This is explained by the complexity of voice data. Beyond the primary identification attributes, the voice encompasses an extensive amount of data, often known as soft biometrics [21]. These are anatomical or behavioral characteristic such as age, gender, ethnicity, and accent [22]. These soft biometrics can be detected automatically via machine learning (ML) systems [23–26] and their integration alongside primary biometrics enhances the precision of the recognition process [27, 28]. In addition, short voice recordings have been used to reconstruct average-looking facial images capturing age, gender, and ethnicity characteristics [29]. Despite their legitimate utility, soft biometrics also usher in possibilities of misuse that can put individuals at risk of privacy concerns without their awareness. This can manifest in unauthorized data processing for illegitimate purposes such as discrimination, invasive advertising, extortion, and other forms of abuse. Moreover, the biometric data of speakers is often stored in external databases, which may not have strict protections. This lack of rigorous security for stored data can lead to breaches and illegitimate access.

Another issue related to ASV is that of concerns regarding biases in ASV systems. Disparities in ASV responses have been noted in the form of differential behavior towards different genders, nationalities, and accents [30, 31]. This leads to discrimination between individuals which is prohibited by the GDPR under the fairness principle. Bias can originate from various sources, such as bias in the data used to train the system, including unbalanced datasets that underrepresent certain groups or biased labeling due to societal biases in human decision-making. The model itself can also exhibit bias, as a machine learning algorithm may prioritize achieving higher accuracy on overall samples at the expense of sacrificing performance on minority groups, thereby placing these minority groups at a disadvantageous position.

1.4 Research Contributions

This dissertation enhances compliance with GDPR principles concerning data privacy and fairness in voice biometrics applications extends beyond the realm of ASV

[5] www.prima-itn.eu/blog/a-reflection-on-privacy-security-and-anonymity

systems to include spoofing countermeasures (CMs) systems. These CMs play a pivotal role in fortifying ASV systems against spoofing attacks, ensuring the integrity of the authentication process (Section 3.1). Our approaches use advanced cryptographic primitives, data perturbation, and machine learning techniques.

Previous studies only considered preserving privacy of ASV systems. However, given the recent proliferation of deepfakes [32], such as voice conversion (VC) [33] and text-to-speech (TTS) [34] technologies, modern ASV systems incorporate CMs that leverage voice characteristics. These CMs can be cloud-based, therefore, are susceptible to privacy breaches.

Recognizing the existing gap in the literature, we introduce PRIVASP, the first solution that ensures not only privacy of individuals but also safeguards intellectual property (IP) of a cloud-based spoofing detection system within real-time applications. To ensure the compatibility of our spoofing CM with secure multi-party computation (MPC), we design a shallow NN model from scratch, adhering to the GDPR privacy-by-design principle.

This work was published in:

- **Oubaïda Chouchane**, Baptiste Brossier, Jorge Esteban Gamboa Gamboa, Thomas Lardy, Hemlata Tak, Orhan Ermis, Madhu Kamble, Jose Patino, Nicholas Evans, Melek Önen, Massimiliano Todisco, "**Privacy-preserving voice anti-spoofing using secure multi-party computation**," in *Proc. INTERSPEECH 2021*, Brno, Czech Republic, September 2021.

In harmony with the principle of data privacy by design of the GDPR, we presented an advanced privacy-preservation data obfuscation technique based on differential privacy (DP) mechanisms. DP mechanisms traditionally serves as mechanisms for data anonymization. We included it to an adversarial auto-ancoder (AAE) model to maintain individual identity while masking gender details. Our approach not only ensures potent gender concealment but also fortifies differential privacy assurances, presenting a flexible balance between preserving privacy and retaining utility of speaker verification.

This work was published in:

- **Oubaïda Chouchane**, Michele Panariello, Oualid Zari, Ismet Kerenciler, Imen Chihaoui, Massimiliano Todisco, Melek Önen, "**Differentially Pri-**

vate Adversarial Auto-Encoder to Protect Gender in Voice Biometrics," in *Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security 2023*, Chicago, United States, June 2023.

In compliance with GDPR principles aimed at ensuring the right of non-discrimination and fairness in automated decision-making systems, we developed an approach that carefully balances trade-offs between speaker verification accuracy, gender data privacy, and fairness. This is executed by fine-tuning the wav2vec 2.0 [35] pre-trained model. While there has been previous work that addresses these challenges, they have typically been handled individually. Our research is the first to address these three aspects concurrently in the context of voice biometrics. Additionally, our study introduces the fairness activation discrepancy (FAD) metric tailored for speech data as a method for analyzing network fairness. This metric is adapted from the InsideBias [36] metric initially designed for face biometrics. It represents a novel application to the voice biometrics domain in our research.

This work was published in:

- **Oubaïda Chouchane**, Michele Panariello, Chiara Galdi, Massimiliano Todisco, Nicholas Evans, "**Fairness and Privacy in Voice Biometrics:A Study of Gender Influences Using wav2vec 2.0**," in *BIOSIG 2023*, Darmstadt, Germany, September 2023.

The absence of an international standard for measuring fairness in biometric systems presents a significant challenge in the field. Besides, most of the scientific research addressing the measurement tools of fairness in biometric recognition systems focus on face recognition (FR). We contributed to the field by assessing three fairness metrics, proposed for the FR field, in the context of ASV. We evaluated their effectiveness in meeting certain criteria set in the literature for selecting metrics for fairness assessment. We then used the most suitable metric to evaluate five state-of-the-art ASV systems and studied the trade-off between utility and fairness. This work serves as a benchmark for future developments of more equitable and effective ASV systems.

This work was submitted in:

- **Oubaïda Chouchane**, Christoph Busch, Chiara Galdi, Nicholas Evans, and Massimiliano Todisco, "**A Comparison of Differential Performance Metrics for the Evaluation of Automatic Speaker Verification Fairness**," in *Odyssey 2024*, Québec, Canada, June 2024.

1.5 Thesis Outline

This section outlines the structure and content of each chapter in the thesis. The remainder of this dissertation is structured as follows: In chapter 2, we provide a background and literature review on privacy and fairness enhancing technologies related to this thesis. We present cryptographic privacy preservation techniques namely secure multi-party computation and homomorphic encryption, as well as differential privacy. We further introduce machine learning techniques studied on our solutions namely disentanglement and adversarial training. Additionally, we present a literature review of fairness assessment in the context of biometrics.

In chapter 3, we identify challenges of preserving privacy of speakers while maintaining utility of the spoofing countermeasures of automatic speaker verification systems. We then present our solution for privacy-preserving voice anti-spoofing using the secure multi-party computation (MPC) technique.

In chapter 4, we delve into the challenges of protecting gender in speaker verification applications. We present our novel solution based on combining differential privacy mechanism and adversarial training to retain identity of speakers and hiding their gender-related information.

In chapter 5, we explore the complexities of ensuring privacy-preserved speaker verification while also evaluating the fairness of our proposed solution. We provide an analysis of how gender information impacts data privacy and fairness within a speaker verification application. Furthermore, we introduce a modified fairness metric for voice biometrics, adapted from facial biometrics.

In chapter 6, we identify the challenge of non-existent international standards for fairness evaluation in the field of biometrics. We assess three fairness measures in the context of speaker verification context and evaluate ASV systems using the most suitable fairness metric.

Finally, in chapter 7, we draw the dissertation to a close by summarizing our results and suggesting potential directions for future research.

Chapter 2

Background and Literature Review

In this section we start by expanding our focus to discuss privacy enhancing technologies (PETs) highlighting their role in safeguarding individual privacy. This exploration forms the basis for the methodologies central to our study, including cryptographic methods, data perturbation techniques, disentanglement strategies and machine learning (ML) approaches. We also delve into the concept of fairness in relation to biometrics. This aspect stresses the importance of equitable treatment and non discriminatory practices in technological applications while recognizing that fairness enhancing technologies (FETs) are still developing compared to established privacy measures. In addition to ethical considerations the section provides a thorough review of existing literature examining past research and contributions in these areas to foster a comprehensive understanding of both well established and emerging technologies, for promoting ethical use and implementation.

2.1 Privacy-Enhancing Technologies

Several technologies have been developed to ensure data privacy. Some of these techniques have been exploited to preserve privacy of voice biometric data. In this chapter, we introduce the building blocks of these solutions that are based on advanced cryptography, data perturbation, and machine learning.

2.1.1 Advanced Cryptographic Techniques

Advanced cryptographic techniques focus on safeguarding data by making it appear random, either by secretly sharing it, using primitives like secure multi-party computation, or by encrypting it using a private key, as in the case of homomorphic encryption, while still enabling the processing of the data. In the upcoming subsections, we will discuss key concepts of these techniques and examine related work in the voice biometrics field.

2.1.1.1 Secure Multi-Party Computation

Secure multi-party computation (MPC) is an advanced cryptographic technique that allows multiple parties to collaboratively compute a function over their private data while keeping them concealed from each other. MPC was first proposed by Andrew Yao in the early 1980s [37] as a two-party scenario. It was demonstrated by the *Millionaire's Problem*, a conceptual challenge where two millionaires aim to ascertain who is wealthier without revealing their actual wealth. Secure two-party computation (2PC) techniques were employed in this scenario to reveal only the identity of the wealthier individual while keeping all other information private. Building upon this foundation, Goldreich et al. [38] later expanded the scope of MPC to include an arbitrary number of parties.

In the context of MPC protocols, an adversary is not only one of the players/parties engaged in the computation procedure. According to the work of Lindell [39], there are primarily three categories of adversaries in an MPC setting. Each adversary represents a different level of threat based on their potential actions and intentions:

- **Semi-honest adversaries (honest-but-curious/passive):** adhere to protocol specifications but seek to learn private inputs of other parties. Protocols secure against such adversaries ensure basic privacy that prevent data leakage.

- **Malicious adversaries (active):** pose a higher threat that the latter by arbitrarily deviating from the protocol for malicious purposes. Protocols secure against active adversaries provide robust protection against adversarial attacks.

- **Covert Adversaries:** combine traits of semi-honest and malicious adversaries. They take the risk of detection when attempting to break the protocol, with a specific probability of being caught.

To enable parties to jointly and privately perform operations, a range of techniques have been proposed in the literature. Prominent among these are Yao's garbled circuits (GC) [40] and secret sharing methods, both additive and Boolean [38, 41]. In this thesis, in Chapter 3, we use a semi-honest 2PC protocol that makes use of additive secret sharing.

2.1.1.1.1 Yao's Garbled Circuits

Yao's garbled circuit is a 2PC protocol that allows two semi-honest parties to jointly evaluate a function f over their private inputs without revealing them to one another. In a Yao's GC protocol one party plays the role of a *garbler* and the other one plays the role of an *evaluator*. Let us consider two parties P_1, the garbler, and P_2, the evaluator, each holding a secret input x and y, respectively. The *garbler* is responsible for converting the function $f(x, y)$ into a *Boolean circuit* that is composed of gates such as AND and XOR. Then, P_1 garbles this circuit by creating garbled gates for every gate in the circuit. An example of function f of an AND gate is presented in Table 2.1.

Input wires		Output wire
x	y	z
0	0	0
0	1	0
1	0	0
1	1	1

Table 2.1: Truth table of an AND gate.

To create the garbled AND gate, P_1 first assigns a uniformly random label (i.e. key) to each bit values 0 and 1 of each wire x, y, and z of the circuit (six keys in total). Then, P_1 encrypts the output keys, $k^z_{(0,1)}$ using the input keys $k^{(x,y)}_{(0,1)}$ as presented in Table 2.2. E_k represents the symmetric encryption procedure using

the key k to ensure that only with the correct keys P_2 can later decrypt the corresponding output z.

Input wires		Output wire
x	y	z
k_0^x	k_0^y	$(E_{k_0^x}(E_{k_0^y}(k_0^z))$
k_0^x	k_1^y	$(E_{k_0^x}(E_{k_1^y}(k_0^z))$
k_1^x	k_0^y	$(E_{k_1^x}(E_{k_0^y}(k_0^z))$
k_1^x	k_1^y	$(E_{k_1^x}(E_{k_1^y}(k_1^z))$

Table 2.2: Encrypted truth table of AND gate.

After randomly shuffling the order of the four output ciphertexts, P_1 shares both the garbled outputs and the mapping from these outputs $(k^z_{(0,1)})$ to their corresponding actual bit values with the evaluator. In order to evaluate the GC, the *evaluator* needs to receive the corresponding keys as well to decrypt the output z of $f(x, y)$. Since the input of the *garbler* is encrypted, sharing the garbled input GI(x) (k_0^x or k_1^x) of the first party reveals nothing about its original input. However, the *garbler* needs to send one key $k^y_{(0,1)}$ to the *evaluator* since sending both keys k_0^y and k_1^y could leak information about the input of the garbler. The *evaluator* also cannot share with the *garbler* its input bit. To receive its GI(y) without revealing any information about its input, both parties make use of a cryptographic technique called Oblivious Transfer (OT) [42]. Finally, P_2 evaluates the function $f(x, y)$ using GI(x) his obliviously obtained key GI(y). To reveal the output, which is the actual result of $f(x, y)$, P_2 decrypts the final output z by correlating the garbled output label $(k^z_{(0,1)})$ with its corresponding actual bit value, using the previously shared mapping from P_1.

2.1.1.1.2 Additive Secret Sharing

In a 2PC setting using additive secret sharing, a secret x is divided into two shares in such a way that these shares sum up to the original secret under modulus arithmetics. Specifically, x is an integer split into two shares: $\langle x \rangle_1$ and $\langle x \rangle_2$. These shares are distributed among two non-colluding parties, P_1 and P_2. The shares are defined as follows:

$$\langle x \rangle_1 = \text{random value} \in \mathbb{Z}_N \tag{2.1}$$

$$\langle x \rangle_2 = x - \langle x \rangle_1 \mod n \tag{2.2}$$

where n is a large integer, typically chosen as a prime number. This methodology operates only with integers in $\mathbb{Z}_N$ and requires the conversion of floating-point numbers into integers for accurate reconstruction under modular constraints. After dividing the secret, each party P_1 and P_2 can independently perform computations on their respective shares.

Computing Addition

To compute the addition of two secrets x and y, each party adds their respective shares of x and y. Let the shares of x be $\langle x \rangle_1$ and $\langle x \rangle_2$, and the shares of y be $\langle y \rangle_1$ and $\langle y \rangle_2$. The sum $x + y$ is computed by combining these partial sums, as illustrated in Figure 2.1.

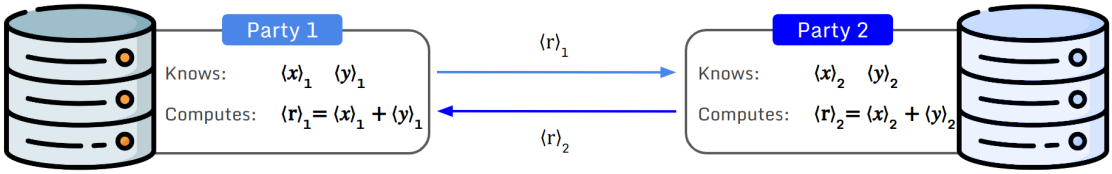


Figure 2.1: Illustration of secure addition in 2PC using additive secret sharing

Computing Multiplication

Multiplication in a 2PC setting using additive secret sharing is more complex and requires an additional interaction round. It is split into two phases: the offline (or preprocessing) phase and the online phase. During the offline phase, a *crypto-provider* supplies additional secret values to the two non-colluding parties to assist in securely computing the multiplication without revealing their private inputs. This technique, illustrated in Figure 2.2, is known as the Beaver Triplets technique [41].

In this method, the crypto-provider randomly selects two values, a and b, and computes $c = a \times b$. The values a, b, and c are then secretly shared between the two parties. Consequently, P_1 receives shares $\langle a \rangle_1$, $\langle b \rangle_1$, and $\langle c \rangle_1$, in addition to its original shares $\langle x \rangle_1$ and $\langle y \rangle_1$. P_2 receives $\langle a \rangle_2$, $\langle b \rangle_2$, and $\langle c \rangle_2$, along with $\langle x \rangle_2$ and $\langle y \rangle_2$.

To compute the product $x \times y$ using the Beaver triplets, each party first performs specific calculations. P_1 and P_2 each compute the differences between their shares of x and a, and y and b respectively. These differences, represented as the shares of A and B in Figure 2.2, are then revealed to the other party. Next, each party computes the differences $x - a$ and $y - b$. These computations lead to the values of z_1 and z_2, as illustrated in Figure 2.2. Subsequently, each party shares its respective shares of z with the other party. The multiplication results of x and y is then obtained by summing these shares z_1 and z_2.

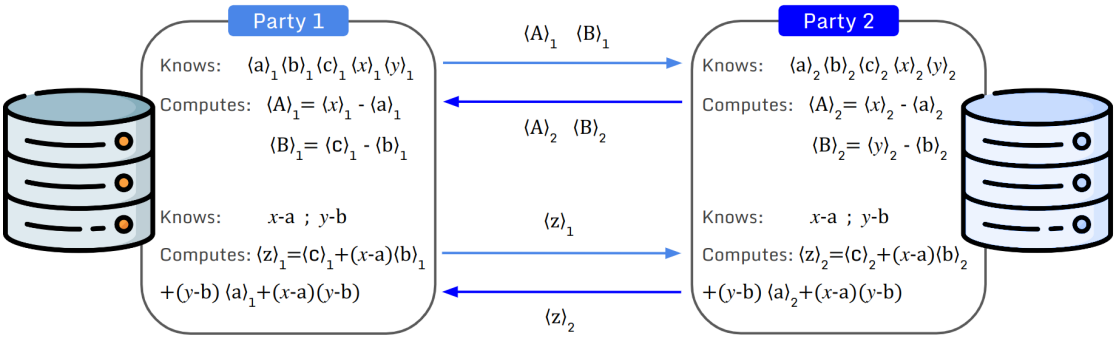


Figure 2.2: Illustration of the Beaver triplets technique for secure multiplication in 2PC

In this thesis

2.1.1.2 MPC-based Approaches in Voice Biometrics

Various researches have contributed to the development of privacy-preserving speaker verification systems using MPC-based solutions. Portêlo et al. [43] propose a privacy preserving Gaussian mixture model-universal background model (GMM-UBM) speaker verification system using Yao's Garbled Circuit (GC). The GMM-UBM system uses a large set of background speakers to train a UBM model, which is a speaker independent-GMM, and adapted user-specific GMMs for each enrolled speaker. During verification, the similarity score is computed by the ratio

of the log likelihoods of an MFCC feature vector x with the user-adapted GMM and the UBM. The score computation function is composed of scalar products and Logsum. In their setting, the authors assume that the two parties involved in the Yao's GC protocol are the user as the garbler and the service provider (i.e. ASV system provider) as the evaluator. The user is responsible for converting the scalar products and logsum operation into a Boolean GC. The service provider is responsible for evaluating the GC and making the authentication decision. Experimental results show that the utility of the privacy-preserving ASV system is almost the same as the non-private system. The execution time of the private GMM-UBM system is efficient, but it increases linearly with the number of GMM components. Another major drawback of this approach is that the service provider has access to the enrollment user-specific GMM model in plaintext which violates the privacy of the speaker. The parameters of the adapted model embody the characterization of the voice of the user.

Aliasgari et al. [44] also make use MPC to preserve privacy of speaker in hidden Markov model (HMM) framework. The authors perform private computations on HMMs in both the semi-honest and malicious models of MPC using secret sharing. The proposed solution is based on floating-point arithmetics and ensures that the HMMs, which contain speaker information, are not available in cleartext to the servers using them to perform authentication.

Additionally, to perform speaker verification while preserving the privacy of users, Treiber et al. [45] propose a semi-honest 2PC approach based on secret sharing and Yao's GC . In this setting, a client arithmetically shares his speech feature vector among two non-colluding servers. The servers then securely compute the similarity score. In order to perform the comparison, the servers switch from arithmetic to Yao sharing. The servers evaluate the garbled function greater than gate $>$ on the Yao-shared similarity score and a pre-fined threshold.

In an attempt to be compliant with the GDPR's requirements of data privacy, researchers made use of MPC to preserve the privacy of speakers. Nautsch et al. [46] introduce a privacy preserving semi-honest 2PC-based solution combined with a homomorphic encryption scheme (see Section 2.1.1.3) for an i-vector based ASV system with cohort score normalisation using PLDA comparisons. The authors propose a cohort pruning scheme which enables efficient selection of the top-n relevant cohort comparisons. This approach operates with binary voice represen-

tations to reduce the computation time for biometric comparisons caused by the homomorphic encryption computations.

In [47], Teixeira et al. make use of MPC in order to design a private feature extraction system for speaker recognition applications. The authors implement a privacy-enhancing secret sharing MPC-based solution with the intention of privately extracting the speaker features while keeping both the speaker voice and the extraction model private. In this work, four cases are considered. In the first, only the client and the model provider are involved in the computations and a semi-honest 2PC protocol is employed. In the second, three semi-honest parties are involved in the computation of the feature extraction: the client, the vendor, and a trusted non-colluding server. In the third scenario, a second semi-honest MPC server is added. This setting allows the assumption of one malicious party: either the client or the ASV system provider. If either parties behave maliciously, the protocol will detect the malicious behavior and will abort. In the fourth scenario, only the client and the vendor run the 2PC protocol, and one of them might deviate from the protocol description and behave maliciously. While this solution provides the strongest level of security, it also incurs the greatest computational and communication expenses.

2.1.1.3 Homomorphic Encryption

Homomorphic encryption (HE), initially envisioned by Rivest, Adleman, and Dertouzos in 1978 [48] and further developed by Gentry in 2009 [49, 50], is a cryptographic primitive that enables computation directly on encrypted data. It allows a third party, like a cloud server, to conduct additive and multiplicative operations on ciphertexts without decrypting them, ensuring that only the data owner can access the plaintext results. At its core, HE operates on the principle of homomorphism, an algebraic property that facilitates operations on ciphertexts as follows:

$$E(m_1 \diamond m_2) = E(m_1) \diamond E(m_2) \tag{2.3}$$

where E denotes the encryption function, and $\diamond$ indicates an operation (i.e. addition or multiplication). This process ensures that the operations on ciphertexts yield encrypted results which, when decrypted, equate to the results of the same operations performed on the plaintexts m_1 and m_2.

Homomorphic encryption categorizes into three principal categories: partially homomorphic encryption (PHE), somewhat homomorphic encryption (SHE), and fully homomorphic encryption (FHE). Each category is defined by the range and the number of permissible operations.

- **PHE:** is capable of performing an unlimited number of either additive or multiplicative operations. PHE [51] includes foundational schemes such as RSA [52] and El-Gamal [53] for multiplication and Paillier [54] for addition.
- **SHE:** supports a limited combination of additive and multiplicative operations [55], as illustrated by the Boneh-Goh-Nissim (BGN) scheme [56], which allows numerous additions but restricts to a single multiplication, and the Polly Cracker scheme [57], notable for its unlimited operational capacity at the expense of ciphertext scalability.
- **FHE:** permits limitless operations, yet repeated homomorphic operations, especially multiplications, amplify errors [58], potentially leading to undecipherable ciphertexts. Gentry's bootstrapping technique [59]] firstly mitigates this error accumulation. The Brakerski-Gentry-Vaikuntanathan (BGV) [60] and Brakerski/Fan-Vercauteren (BFV) [61,62] cryptosystems are prominent examples leveraging the Learning With Errors (LWE) or Ring Learning With Errors (RLWE) problems for their security foundation. The Cheon-Kim-Kim-Song (CKKS) FHE scheme [63] is recognized for its ability to handle approximate arithmetic on ciphertexts with real or complex number vectors and has relatively compact keys and small ciphertexts, which makes it more efficient and practical than some other FHE schemes.

Developing an HE scheme typically involves four stages: key generation, encryption, execution of homomorphic operations (i.e. evaluation), and decryption. Key generation creates cryptographic keys, encryption transforms a message into ciphertext, evaluation executes a function over the ciphertexts, and decryption recovers the original message from the ciphertext.
HE schemes are classified as either symmetric or asymmetric. In the symmetric HE schemes the same key is used for both encryption and decryption which requires secure key sharing. On the other hand, asymmetric HE schemes use a public key for encryption and a private key for decryption which enhances security by

eliminating the need for key exchange.

In the context of asymmetric HE system, a client encrypts sensitive data (m_1 and m_2) using a public key pk, visually highlighted in yellow in Figure 2.3 and outsources computations to an untrusted third party, like a cloud server. This encrypted data is denoted as $E(m_1)$ and $E(m_2)$ in Figure 2.3 to indicate their encrypted state. The server processes these encrypted inputs by applying a specified function, symbolized by a $\diamond$, that delivers an encrypted outcome. This result is then sent back to the client, who can decrypt it using its private key sk (represented in gray in Figure 2.3), ensuring that only the client can access the plaintext result of the computation.

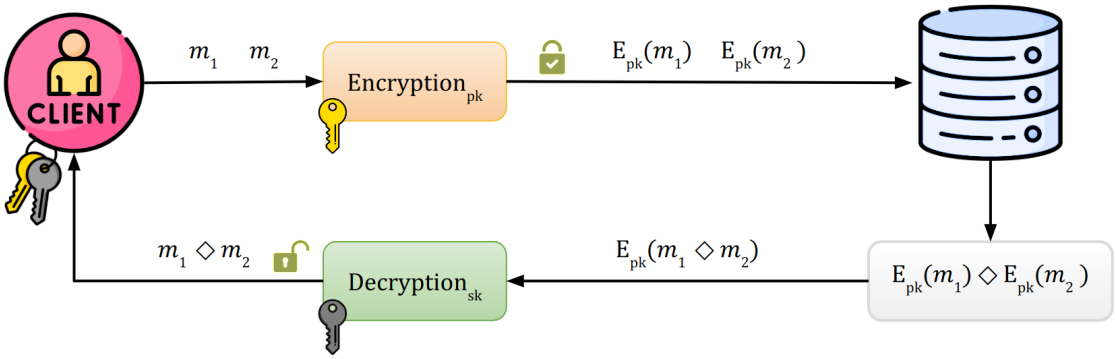


Figure 2.3: Illustration of asymmetric homomorphic encryption

2.1.1.4 HE Application to Voice Biometrics

HE-based methods have been used to preserve privacy of speaker in ASV systems. The work in [64–66] introduces a privacy-preserving protocols for ASV systems based on GMM using the PHE Paillier scheme [54] in combination with a semi-honest 2PC protocol. During enrolment, the user generates a private key and a public key to share it with the ASV system provider. In order to generate the user-specific GMM, the system provider sends the UBM in plaintext to the user. The user then encrypts the GMM with its private key and sends it to the system provider. This technique guarantees the privacy of users during the enrolment phase. During verification, the user encrypts the MFCC features with its private key and send it to the ASV system provider. The system then computes the log-likelihoods in the encrypted domain for the UBM and the components of the encrypted GMM of the claimed user. The similarity score is derived from

component-wise encrypted log-likelihoods using the logsum protocol. This operation requires additional communication between the user and the ASV service provider. This approach ensures that the features of the user are kept private from the system, and the models stored in the system database are kept hidden from the user. Although this privacy preserving solution maintains verification capabilities of the baseline system, it leads to a significant computational overhead compared to the GMM-UBM baseline ASV system. This is caused by the large amount of time required to perform operations in the encrypted domain.

In order to be compliant with the data privacy principle of the GDPR, Nautsch et al. [67] also propose Paillier-based approach to preserve privacy of speakers in an i-vector-based ASV using cosine or PLDA as similarity measure. This approach not only protects the speech data of the user, but also the model provided by an AS_{vendor} during verification. In this work, the user in assumed to interact with two non-colluding servers $DB_{controller}$ and $AS_{operator}$ to privately perform the authentication task. The $AS_{operator}$ generates a pair of public and private keys (pk and sk) and shares pk with the user. During enrolment, the user extract the reference feature vector, encrypts it using pk, and sends it to the $DB_{controller}$. During verification, the user extracts and encrypts their probe feature vector and requests the encrypted reference vector previously saved in $DB_{controller}$. Upon reception, the user homomorphically computes the similarity over the encrypted inputs and obtains an encrypted score. This score is then transmitted to the $AS_{operator}$, which decrypts it using the private key sk and compares it to a pre-fixed threshold. Based on this comparison, the $AS_{operator}$ either accepts or rejects the claimed identity. Experiments prove that this solution maintains the verification performance while meeting privacy requirements. However, the use of HE leads to a significant increase in communication and computation overhead. This renders the solution impractical, especially for computationally limited devices like mobile phones. Moreover, this approach is vulnerable in terms of security considering a malicious client who can cheat the system by sending an encryption of an accepting score to the $AS_{operator}$.

To preserve the privacy of speakers as mandated by the GDPR, the author in [68] designs FHE-based scheme to compute the cosine similarity in the encrypted domain. During enrolment, the user locally extracts the speech embedding (i.e. features), and generates a pair of public and private keys (pk and sk) for encryption

and decryption. The user then applies the CKKS FHE [63] scheme to encrypt its embedding vector using the public key pk and sends it along with its ID and the public key to the server to get enrolled. During the verification phase, the speaker follows the same steps and encrypts its embedding vector using the same pk used during the enrolment. The server receives the encrypted test vector as well as the user ID which is used to search for the reference vector saved in the database. Next, the server evaluates the cosine similarity in the encrypted domain. After generating the encrypted score, the server sends it to the user. Finally, the user's device decrypts the score and determines whether it is above the threshold needed to authenticate the user. This final step poses a security threat as a malicious user can modify the received score to gain illegitimate access.

2.1.2 Differential Privacy

The question of *how to analyse data while preserving privacy of individuals* has puzzled researchers for decades, especially considering that data cannot remain useful if it is fully private. In 2006, Cynthia Dwork and her colleagues [69, 70] successfully addresses this riddle and introduced a concept that strikes a balance between privacy and utility, as well as quantifies privacy loss. This concept is known as Differential Privacy (DP) [71], also referred to as global differential privacy (GDP) or centralized DP. The main idea behind DP is that when querying a database for analysis, the outcome of these queries should not disclose the participation of any specific individual in the database. This privacy promise is effective even against an adversary (i.e. an individual or organization attempting to extract sensitive information about people from data analysis results) who possesses unlimited computational power and comprehensive knowledge of both the DP privacy-preserving approach and the system used for gathering and processing the data. DP also promises individuals that no additional harm will arise from their data being in the database, a harm that would not exist if their data were absent.

It is crucial to understand that DP itself is not an algorithm but a *definition* or a framework. For a given computational task T, numerous differentially private algorithms (also known as *mechanisms*) can be designed to execute T in an ϵ-differentially private way. $\epsilon > 0$ is known as the *privacy budget* that provides a measure of the privacy loss incurred by the DP algorithm. The smaller the

value of ϵ, the smaller the privacy loss (i.e. the stronger the privacy protection) and vice versa. The DP mechanisms can vary in their accuracy and effectiveness. For smaller ϵ values, it become more challenging to design a highly accurate DP algorithm.

2.1.2.1 Local vs Global Differential Privacy

Various mechanisms are employed in DP to protect data privacy. This technique

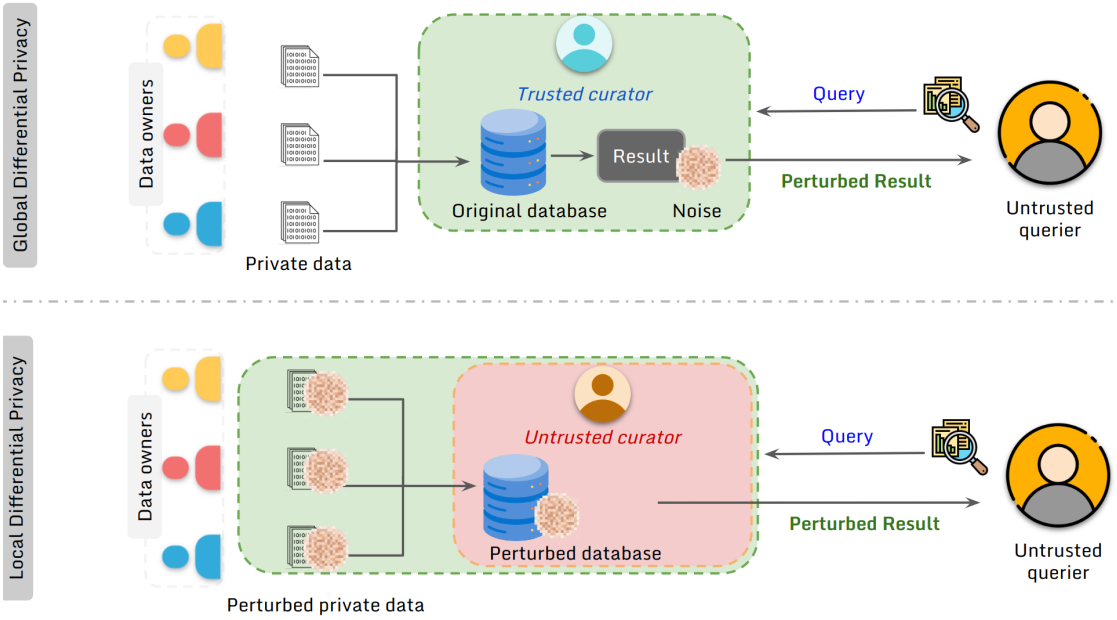


Figure 2.4: Global vs. Local Differential Privacy

involves integrating carefully calibrated noise into the dataset, which can be applied either to the query results or directly to the data inputs. Based on the noise injection point, DP is classified into two categories: global differential privacy and local differential privacy (LDP), as depicted in Figure 2.4.

GDP relies on a trustworthy central authority, referred to as the *data curator*, to gather individual data from data owners and synthesize it. When an untrusted data querier requests information from this collective database, the curator introduces carefully calibrated noise derived from a DP mechanism to the output of the query prior to its release. The querier then receives the perturbed result that maintains meaningful information while preserving the privacy of individuals present in the database. However, the main drawback of GDP is the inherent need

to place trust in a central entity, which may not always be feasible in practical, real-world scenarios.
LDP overcomes this limitation by enabling data owners to add noise to their own data independently, thus eliminating the need for a central trusted authority. LDP consists in protecting individual input data before its collection, ensuring that the privacy of each user is locally preserved (i.e. at the source). This entails safeguarding the privacy of each individual record at its source, rather than applying protection to the dataset as a whole post-aggregation.

In this thesis, we make use of an LDP mechanism in Chapter 4. Below is a formal definition of LDP.

2.1.2.2 Local Differential Privacy

Local differential privacy plays a crucial role in protecting personal data and assessing the privacy risks. In this paragraph, we formally introduce the LDP model and give a brief overview of the related notions.

(ϵ)-local differential privacy is defined as follows:

Definition 1 (Local Differential Privacy [72]) *A randomized algorithm $\mathcal{M}$ satisfies (ϵ)-LDP if and only if for any pairs of input values $x, x' \in \mathcal{X}$ in the domain of $\mathcal{M}$, and for all possible outputs $S \subseteq Range(\mathcal{M})$, we have:*

$$Pr[\mathcal{M}(x) \in S] \leq e^{\epsilon} \cdot Pr[\mathcal{M}(x') \in S] \tag{2.4}$$

where Pr denotes the probability. This definition quantifies the level of privacy protection provided by $\mathcal{M}$ and ensures that the likelihood of observing a particular output is not significantly affected by the choice of input between x and x'.

ℓ_1-Sensitivity. The ℓ_1-sensitivity [73], denoted as Δf, is a measure of the maximum influence that a single data point can have on the result of a numeric query f. In an LDP mechanism, the ℓ_1-sensitivity can be defined as shown in (2.5), where x and x' represent two adjacent records in a dataset $\mathcal{X}$, meaning they differ only in one data point, and $\|.\|$ denotes the ℓ_1 norm of a vector.

$$\Delta f = \max_{x,x' \in \mathcal{X}} \|f(x) - f(x')\|_1 \tag{2.5}$$

The ℓ_1-sensitivity is the maximum difference between two adjacent records in a dataset and it provides an upper bound on the potential impact of an individual

record, essentially measuring the worst-case scenario impact of a single individual's record on the function's outcome. The ℓ_1-sensitivity defines the magnitude of the noise needed in order to meet the (ϵ)-LDP requirements.

Laplace Mechanism. The Laplace mechanism [74] is a widely adopted technique for achieving (ϵ)-DP. The Laplace mechanism, presented in (2.6), is initially proposed and adopted for (ϵ)-GDP and can be adapted for (ϵ)-LDP. The mechanism works by adding random noise, sampled from the Laplace distribution centered at 0, to the output of a function in order to obscure any sensitive information about individual records in the database. The amount of noise added is determined by the sensitivity Δf of the function and the privacy budget ϵ. The more sensitive the query is, the more noise is needed to achieve a stricter privacy guarantee. Formally, given a database $\mathcal{X}$ and a function $f : \mathcal{X} \to \mathbb{R}^d$ that maps the database to d real numbers, the Laplace mechanism is defined as:

$$\mathcal{M}(f(x), \epsilon) = f(x) + (n_1, n_2, ..., n_d). \tag{2.6}$$

where each $n_i \sim \text{Laplace}\,(\Delta f/\epsilon)$ is drawn from the zero-centered Laplace distribution with scale $\Delta f/\epsilon$.

The Laplace mechanism has been demonstrated to be particularly effective in the context of numerical queries (e.g. counting queries, histogram queries, and classification queries) with low sensitivity [71]. In Chapter 4, we use the Laplace mechanism in the context of a gender classification task.

Key Properties of Differential Privacy Local differential privacy possesses three main properties: sequential composition, parallel composition, and post-processing [71, 74, 75].

The first property, *sequential composition*, is delineated as:

Definition 2 (Sequential Composition) *Consider a sequence of n mechanisms $\{M_1, \ldots, M_n\}$, each providing ε_i-LDP. When these mechanisms are applied in sequence to a dataset, the resultant composite mechanism, denoted by $(M_1, \ldots, M_n)$, adheres to a cumulative privacy guarantee expressed as $(\sum_{i=1}^{n} \varepsilon_i)$-LDP.*

The second key property in DP , called *parallel composition*, provides an alternative approach to assess the privacy cost of multiple data releases. It involves partitioning the dataset into disjoint segments and applying a differentially private mechanism separately to each segment. Formally,

Definition 3 (Parallel Composition) *Assume a set of n mechanisms* $\{M_1, \ldots, M_n\}$*, with each ensuring* ε_i*-LDP independently. When operated on disjoint subsets of the dataset, the combined effect of these mechanisms, represented by* $(M_1(D_1), \ldots, M_n(D_n))$*, conforms to a privacy level characterized by* $(\max(\varepsilon_i))$*-LDP.*

The *sequential composition* and the *parallel composition* properties are essential to allocate privacy budgets effectively and ensure that the overall privacy of the system is maintained.

The third property is *post-processing*, meaning that any function applied to the output of a differentially private mechanism cannot weaken its privacy guarantees. This immunity ensures that the privacy level is maintained, regardless of any further analysis or transformation performed on the data. Formally, the post-processing is defined as follows:

Definition 4 (Post-processing) *Let* $\mathcal{M}$ *be an* ϵ*-differentially private mechanism and* g *be an arbitrary mapping from the set of possible outputs to an arbitrary set. Then,* $g \circ \mathcal{M}$ *is* ϵ*-differentially private.*

In Chapter 4 of this thesis, the post-processing property ensures that any arbitrary computation performed on the speaker embedding, which is the output of a DP mechanism, does not compromise the privacy guarantees.

2.1.2.3 Differential Privacy in Biometrics

Differentially private solutions were proposed for more than a decade and regarded as a privacy protection tool for different areas [76–79]. Recently, DP mechanisms started to gain the attention of the biometrics field researchers. Chamikara et al. [80] design PEEP (Privacy using EigEnface Perturbation), a privacy preserving approach for face recognition system using local differential privacy. The authors assume that any input device used to capture face images make use of PEEP. The device starts by capturing the face images of individuals. Then, a technique called principle component analysis (PCA) is used to identify patterns in data, reduce the dimensionality of these images while only keeping the most relevant features. This dimensionality reduction technique increases computational efficiency while maintaining high accuracy. In a face dataset $D = x_1, x_2, ..., x_n$, each x_i represent a face vector, and t_i denotes the flattened vector x_i. The first step of PCA is to

standardize D by computing the mean of each flattened face vectors t_i to construct the mean face vector F_m, and subtracting F_m from each element t_i. The second step consists of calculating the covariance matrix C of the standardized data to measure the correlations between the data variables. Next, the third step is referred to as *Eigendecomposition*, which involves the calculation of the eigenvectors and eigenvalues of the covariance matrix. The eigenvectors of C represent the directions along which there is the highest variance, indicating the most significant information. Eigenvalues serve as coefficients associated with eigenvectors to quantifying the amount of variance carried in each eigenvector. The eigenvectors are then sorted in descending order by their corresponding eigenvalues. The eigenvector corresponding to the highest eigenvalue is the first principal component. The final step is to select the first k sorted eigenvectors, where k represents the reduced dimensionality. At this point, the authors propose to add a Laplacian noise to the selected eigenvectors/eigenfaces to randomize the facial representation. This randomized data is shared with an untrusted server which use to train a face recognition model. During the testing phase, the same dimensionality reduction and eigenfaces randomization procedure is performed. The serve only receives perturbed data to carry out the face recognition task. The proposed LDP-based solution guarantees users privacy during both training and testing, and eliminates the need of a trusted party. The accuracy of the face recognition model drops when applying LDP measures, which is expected. Setting a high privacy budget still ensures privacy against reconstruction attacks while maintaining acceptable accuracy.

Shamsabadi et al. [81], propose the Laplace-based differentially private speaker anonymization system. Speaker anonymization is the process of removing the identity related features from the speech content in order to conceal the identity of the speaker while keeping all other aspects intact. The authors suppress the identity of speaker using an x-vector-based speaker anonymization approach. An x-vector [7] is a speaker embedding that represents the identity of a speaker. The first step of the anonymization pipeline is to extract the pitch and bottleneck (BN) features as well as the x-vector from the input speech. Pitch features are the fundamental frequency F_0 of a signal which carry out prosodic information (i.e. intonation, stress and rhythm). BN features are low-dimensional phonetic representation extracted from an intermediate layer of an automatic speech recognition (ASR) model (i.e.

a model that transforms speech into text) and are effective in improving the accuracy of ASR systems [82]. During the second step, the x-vector is anonymized using an external pool of speakers. Finally, the speech waveform is synthesized using the pitch and BN features and the anonymized x-vector. The authors propose a DP pitch and BN extractors where a Laplace noise is included to the architecture of these models. The DP pitch extractor is an auto-encoder model with a Laplace noise added to the output of the encoder. Laplace noise is incorporated into the ASR model to generate differentially private BN features. Experiments showed that this approach guarantee speakers privacy while maintaining high level of utility in term of ASR performance.

2.1.3 Challenges in Applying Advanced Cryptography and Differential Privacy to Machine Learning Based Systems

Nowadays, the foundation of voice biometric systems predominantly relies on machine learning techniques, particularly neural networks. There exist a large number of advanced cryptographic based solutions proposed to build privacy-preserving neural networks [83, 84]. These approaches can be classified into three main categories: (i) MPC-based solutions such as [85, 86]; (ii) HE-based techniques, like [87, 88]; (iii) Hybrid solutions such as [89] and [90] that combine the use of MPC and HE.

However, the integration of MPC and HE primitives to NN-based systems is not straightforward. Computations in the privacy-preserved domain are often expensive, particularly with NNs and especially DNNs, which require numerous linear (addition, multiplication) and non-linear (e.g. activation function, comparisons) operations. In HE schemes, ciphertexts contain noise that grows during homomorphic evaluation operations. Computations will involve larger data than the original plaintext, eventually leading to noise overflow which render the decryption impossible. A technique called bootstrapping [59] was introduced to reduce the noise. However, this approach is costly and significantly increases computational overhead.

On the other hand, MPC protocols show an improvement in execution time compared to HE-based solutions. However, they require multiple rounds of interaction, particularly during multiplication, and involve communications between

parties participating in secure computation. This results in a communication overhead. Additionally, MPC techniques like additive secret sharing are limited to integers, whereas speech data is typically represented as real data. This necessitates converting the data, which may result in information loss and consequently lead to a drop in system accuracy. Non-linear operations as well need to be approximated in order to be compatible with MPC primitives, which adds more challenges.

Last but not least, GDP mechanisms require placing trust in an external third party to manage the data in clear, which poses a privacy threat in real-world scenarios. Additionally, DP mechanisms involve adding noise to data. The more noise added, the greater the privacy guarantee, but at the expense of utility. Particularly in LDP mechanisms, noise is directly added to the data, and during NN operations, the noise is amplified, leading to a drop in data utility. Balancing noise calibration to maintain the trade-off between privacy and utility is challenging.

2.1.4 Disentangled Representations Learning

In machine learning domain, disentangled representations learning (DRL) is a learning paradigm where ML models are structured to acquire representations adept at identifying and disentangling (i.e. separating out) the underlying generative factors of variation embedded within the observed data [91,92]. This definition is based on the concept that the observed data consist of informational factors, where certain factors will exhibit variation, while others will remain invariant. Identifying the generative factors of variation enables learning disentangled representations [92, 93]. In speech domain, the disentanglement relies on the specific informational factors desired and their intended applications [94]. The process of factorization has been used to remove noise factors from speech representations as a speech enhancement technique [95]. In speech synthesis, DRL has been used to independently control different aspects of synthesized speech by disentangling generative factors such as speaker identity, noise level, and speaking rate [96–98]. In speaker recognition, DRL has been used to disentangle speaker identity-related and identity-unrelated information to enhance speaker recognition capabilities by removing irrelevant information [99–102].

DRL approaches can be categorized based on the representation structure into two groups: dimension-wise and vector-wise methods [92].
In dimension-wise methods, a disentangled representation is composed of one or

more dimensions where each dimension represents only one generative factor. For instance, in speech data, each dimension represents one factor such a spitch, volume, and speaking rate. This technique enables precise control over individual aspects of the synthesized data.

In vector-wise methods, a single vector is used to represent one coarse-grained generative factor (i.e. a combination of multiple factors). In speech for example, distinct vectors represent combinations of factors like speaker characteristics, emotional content, and environmental factors. One vector might represent a blend of speaker identity and gender, while another represents emotional content. In this thesis, we introduce a vector-wise DRL technique in Chapter 4

Dimension-wise methods are typically tested on synthetic and simple datasets which often contain numerous fine-grained latent factors. On the other hand, vector-wise methods are commonly employed in real-world scenarios (such as identity swapping, image classification, subject-driven generation, and video understanding) which concentrate on two or more coarse-grained factors [92].

Real-world datasets and applications typically focus on two or several broad factors, such as identity and pose, which align better with vector-wise disentanglement.

2.1.4.1 DRL Techniques

DRL methods mainly make use of generative models like variational auto-encoder (VAE) [103] and generative adversarial networks (GAN) [104].

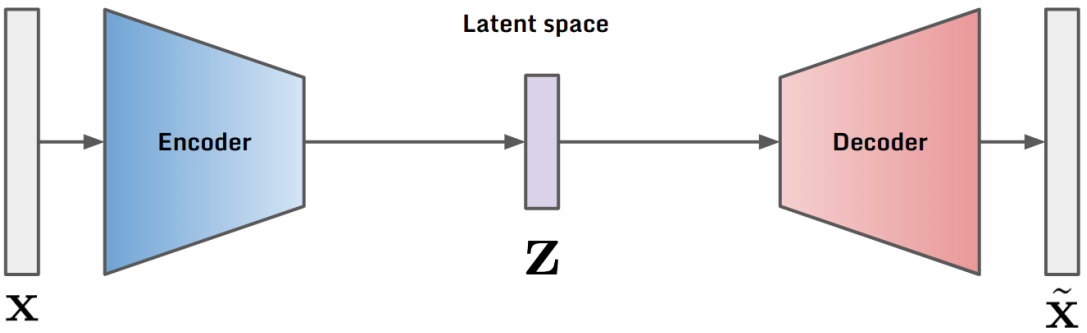


Figure 2.5: Architecture of a variational auto-encoder.

VAE-based DRL techniques, such as those explored in the works by Higgins et al. [105], Burgess et al. [106], Kim et al. [107], Kumar et al. [108], and Chen

et al. [109], harness the capabilities of VAEs to learn rich data representations. Figure 2.5 present the architecure of an VAE. At its core, a variational auto-encoder [103] is a variant of the traditional auto-encoder(AE), a type of neural network that is used to learn representation of an input data. An AE is composed of two NN modules: and encoder and a decoder. The encoder projects an input x from an n-dimensional space to a lower d-dimensional latent space. The decoder then uses the compressed encoded representation to reconstruct the original input x. A loss function is used to quantifies the discrepancy between the original input and its reconstruction. The primary objective of an AE is to uncover the essential features to accurately reconstruct the input data with the fewest possible dimensions. VAE enhances the capabilities of an AE by encoding the input data into a probability distribution over the latent space instead of mapping it to a single point. This probabilistic approach allows VAEs to capture uncertainty and variation in the representation. VAEs enhance disentangled learning by implementing various regularization techniques and loss function modifications during training. These include imposing constraints on the latent space to encourage the separation of underlying factors of variation in the data.

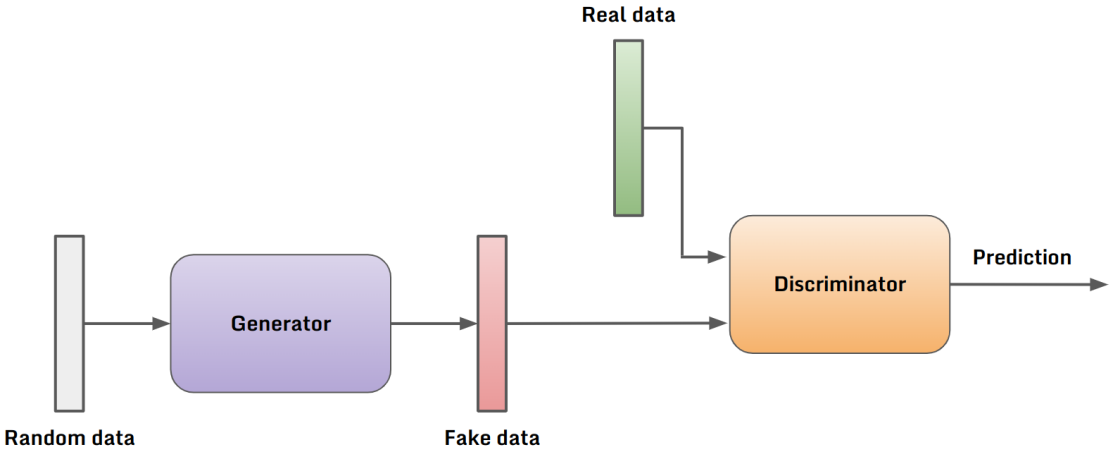


Figure 2.6: Architecture of a generative adversarial network.

Additionally, GAN-based approaches [110–114] provide an alternative path in DRL that enables advanced data generation and manipulation. GAN consists of two primary components: a generative network referred to as the *generator* and a discriminative network known as the *discriminator*, as depicted in Figure 2.6. The role of the generator is to create realistic samples while the discriminator learns

to distinguish between real samples from the dataset and fake samples generated by the generator. During training, the generator and discriminator are trained together in an adversarial manner. The generator learns to produce increasingly realistic samples that are difficult for the discriminator to differentiate from real ones, while the discriminator improves its ability to distinguish between real and fake samples. This adversarial process improves the capabilities of both networks which result in the generation of high-quality samples by the generator. In the ideal scenario, the generator achieves the ability to produce images that closely resemble real ones, causing the discriminator to fail in distinguishing between generated and real images. GANs contribute to disentangled representations learning by modifying network architectures and loss functions. Techniques such as adding auxiliary classifiers or using specific regularization terms enhance disentanglement in the learned latent space of GAN.

In this thesis, combination of an autoencoder adversarial module is used for creating privacy protected speaker embeddings in Chapter 4.

2.1.4.2 DRL in Privacy-Preserving Voice Biometrics

Disentangled representation learning has been utilized as a GDPR-compliant technique for data privacy. Depends on the mainstream task, these techniques can be categorized in two groups: (i)identity anonymization and (ii)soft biometrics preservation.

2.1.4.2.1 DRL for Voice Anonymization

Voice anonymization techniques aim to disentangled the identity of the speaker is disentangled while retaining the spoken content unchanged. This technique is promising in the ASR field where the identity information of the speaker is irrelevant to the required task. Different approaches [115–120] have been proposed for voice anonymization including disentangled representation learning-based techniques [121, 122]. Moreover, The VoicePrivacy 2020 and 2022 challenges (VPC) [123, 124] were introduced to evaluate privacy-preserving ASR modeling frameworks focusing on removing speaker identity information from speech while preserving linguistic content, paralinguistic attributes, intelligibility and naturalness. Privacy is objectively evaluated by assessing speaker verification/re-

identification ability. Baseline systems provided by the challenges are either signal-processing-based [117] or DLR-based. Signal-processing-based systems does not need training data and directly modifies speech characteristics such as the pitch, spectral envelope, and time scaling. The DLR-based baseline system operates in three steps. First, it disentangles frequency information (F_0), bottleneck features, and speaker identity (i.e. x-vector [7]) features from the original audio. Second, the disentangled x-vector is replaced with another dissimilar x-vector drawn from a pool of other users. Finally, an anonymized speech is synthesized using the anonymized x-vector along with the other features. Most systems submitted to the VPCs are inspired by the DLR-based baseline system [1] [125]. Evaluation results of VPC2022 shows that x-vector DLR-based outperform others. In VCPC 2022, Meyer et al. [126] propose a system that achieved the best speaker anonymization preformance. The authors make use of a GAN-based model to extract speaker identity features from the audio.

Thanks to the common datasets, evaluation metrics and baselines systems provided by the VPCs, numerous disentanglement-based voice anonymization systems have been proposed [127–131].

2.1.4.2.2 DRL for Soft Biometrics Privacy Preservation

Another direction for ensuring GDPR compliance to data privacy in the field of biometrics involves disentangling soft biometric attributes while retaining identity-related information for recognition tasks. Several researchers have centered their efforts on developing DLR-based solutions that are capable suppressing soft biometric attributes for biometric data. These techniques are either directly applied to the collected biometric data like face images and voice signals (i.e. at sample level) [132–137] or to the extracted features (i.e. at feature level) [138–141].

Mirjalili et al. [136] introduce a semi-adversarial network (SAN) using an adversarial convolutional auto-encoder (CAE) to conceal gender information in facial images while maintaining their biometric recognition utility. Their approach involves training a CAE in an adversarial manner to produce altered facial images capable of deceiving a discriminator (i.e. gender classifier). During the training, an auxiliary gender classifier ensures the concealment of gender attributes

[1] `https://www.voiceprivacychallenge.org/results-2022/`

within the facial images, while a face matcher verifies the accuracy of face recognition preservation. In a subsequent study [135], the authors present an ensemble of semi-adversarial networks (SANs). This ensemble is constituted of multiple auxiliary gender classifiers and face comparators, which collectively generate diverse perturbations for an input face image. The underlying concept is that among these perturbed images, at least one effectively misleads any arbitrary gender classifier. In their work [137], Mirjalili et al. attempt to enhance the generalization capability of SAN models by combining a variety of face perturbations. While these models effectively preserve the privacy of gender attributes as mentioned previously, their capacity to generalize to arbitrary classifiers is limited. Tang et al. [132] introduce an alternative gender adversarial network model, which effectively conceals gender attributes while maintaining image quality and recognition performance. In addition, this model demonstrates a capacity to generalize across previously unseen gender classifiers. Further work is proposed by Bortolato et al. [140] aimed at enhancing privacy preservation in face images at the template level. The authors propose an AE-based Privacy-Enhancing Face-Representation learning network (PFRNet), that effectively disentangle gender attribute information from identity. This approach results in a good generalization performance across diverse datasets. Furthermore, Terhöst et al. [141] introduce an incremental variable elimination (IVE) algorithm, which trains a set of decision trees to ascertain the importance of variables essential for predicting sensitive attributes. These identified variables are then incrementally removed from facial templates to suppress gender and age features while retaining high face-recognition performance. Building upon this concept, Melzi et al. [138] extend the approach to safeguard multiple soft biometrics, including gender, age, and ethnicity, within facial images.

In speech-related literature, Aloufi et al. [134] build a voice conversion system based on a cycle-GAN architecture. This system is capable of concealing the emotional state of the users while maintaining speech recognition utility. Similarly, in a study by Benaroya et al. [133], a neural VC architecture is introduced to manipulate gender attributes within voice signals. This VC architecture involves multiple auto-encoders designed to convert speech into independent linguistic and extra-linguistic representations. These disentangled representations are learned through an adversarial learning process and can be fine-tuned during voice conversion.

At the template level, Noé et. al. [139] introduce an adversarial auto-encoder

(AAE) architecture aimed at disentangling gender attributes from x-vector speaker embeddings. In their proposed architecture, an external sex classifier is integrated with the AE and attempts to predict the gender from the encoded representations. This AAE system is trained adversarially to mislead the sex classifier. This method effectively conceals gender-related information within the x-vector embeddings while preserving good ASV performance.

2.2 Fairness and Bias Issues

The recent advancements in biometric recognition systems rely on machine learning algorithms, particularly deep learning models [142]. However, ML algorithms, including biometric recognition systems [143, 144], have been shown to be susceptible to biases that impact their decisions to be unfair/biased [145, 146]. ML algorithms can exhibit bias for several reasons, including inherent human bias in the labeling of training datasets, unbalanced datasets that under-represent certain groups, and training mechanisms that prioritize achieving high performance on majority groups at the expense of minority groups. Bias in automated decision-making ML algorithm like biometric recognition systems can significantly impact people's lives when employed in sensitive position like border control [144]. Therefore, it becomes urgent to study fairness of such algorithms.

2.2.1 Definitions

Fairness in the machine learning literature refers to the general idea of treating individuals or groups without bias based on their inherent or acquired characteristics such us age, gender, ethnicity, accent, genetic features, and political opinion [145, 147]. However, there is currently no universally agreed-upon definition of algorithmic fairness. Several fairness definitions have been proposed [148–157]; nonetheless, they are incompatible and cannot be simultaneously employed [158, 159]. Research suggests that fairness in machine learning varies depending on the context [160, 161], which also applies to machine learning-based biometric systems [143]. Therefore, the appropriate fairness definition and metrics of fairness evaluation depends on the application. In this thesis we focus on fairness in voice biometrics verification systems.

2.2.2 Fairness and Bias in Biometrics

Fairness in biometrics system has emerged as a relatively new field of study and has gained significant attention recently. A biometric system or algorithm is deemed biased when noticeable variations in its operation are observed across various demographic groups of people [143]. In an effort to facilitate discussion on algorithmic fairness in biometric systems, Haward et al. [162] introduce two terms: *differential performance* and *differential outcome*. Differential performance refers to the difference in the mated and non-mated score distribution between specific demographic groups for a given biometric task, regardless of any decision threshold. On the other hand, differential outcome deals with differences in false match (FM) (i.e. false acceptance) or false non match (FNM) (i.e. false rejection) error rates among demographic groups relative to a decision threshold.

Most studies evaluating the fairness of biometric systems mainly focus on differential outcome metrics [163–165]. These metrics are easy to calculate, using established error rates, and treat the biometric system as a blackbox [166]. Recently, Kotwal et al. [166] propose an fairness measures based on differential performance to evaluate bias in biometric recognition system. The authors suggest additional measures to complement rather than replace outcome-based fairness measures. Both evaluation approaches work together to analyze the demographic fairness of a biometric verification system. In this thesis, we make use of differential outcome-based metrics (Chapters 5 and 6).

2.2.3 Fairness Assessment for ASV

Researchers have studied demographic biases in various biometric recognition systems, such us facial recognition, fingerprints, palmprints, iris, and finger veins [143]. While face recognition has been the focus of much bias detection and mitigation works over the past decade, there has been limited research on fairness in voice-based biometric recognition systems.

In their study [167], the authors discover that a DNN model shows varying equal error rates (EERs) among individuals based on their language, gender, and age. However, this study only examins one type of NN architecture and a few demographic groups. Building upon this research, Fenu et al. [168] propose a benchmark to assess the fairness of DNN models using two types of DNN-based ASV systems using differences in EER. They train the models with speakers of

different ages, genders, and languages.

In addition, the work presented by Hutiri et al. [169], inspired by Suresh and Guttag's Framework for Understanding Sources of Harm [170], present a comprehensive analysis of bias in the ML development workflow of speaker verification. Suresh and Guttag identify seven sources of bias-related harms across the machine learning life cycle, grouped into two streams: (i) data generation (historical, representational, and measurement bias), and (ii) model building and implementation (learning, aggregation, evaluation, and deployment). These biases originate from: carried out stereotypes, underrepresented groups in the dataset, features and labels of the dataset, overlooking group differences, modeling choices that amplify performance disparities, the use non-representative benchmarks, and the mismatch between deployment and model design. The presented study reveals biases at all stages of development in the VoxCeleb Speaker Recognition Challenge [171], affecting female speakers and non-US nationals the most.

Furthermore, Fenu et al. [172] conduct fairness assessments using three algorithmic fairness definitions and at different operating points (i.e. decision thresholds). They calculate the false acceptance rate (FAR), false rejection rate (FRR), and fairness estimate to explore the trade-off between fairness, security, and usability for three DNN-based ASV systems.

Toussaint and Ding [173] suggest a fairness evaluation framework for ASV systems, using a minimal detection error trade-off (minDET) and DET curves [174] that focuses on a single operational threshold derived from speaker verification scores. However, the proposed DET curves depend on demographic-specific thresholds, making them unsuitable for deployment because of the sensitive and complicated nature of inferring private backgrounds of users.

2.2.4 Bias Mitigation for ASV

Different solution have been proposed in the speaker verification literature in order to mitigate bias therefore improve fairness of ASV systems. These techniques can mainly be categorized in two groups: (i) pre-processing, and (ii) in-processing approaches [175].

Pre-processing methods are based on balancing dataset in terms of demographic characteristics. Fenu et al. [168] make use of this approach and trained ASV models with balanced training data with respect to gender, language, and

age. Experiments show that balancing the dataset led to fairer treatment of groups which reduced the disparity in FAR and FRR between different gender and age demographic groups. However, it does not consistently improve the ability of the models to equally recognize users across all demographic groups. This indicates that simply balancing data is insufficient to achieve fairness in the outcomes of ASV systems. Estevez and Ferrer [176] use a balanced dataset to evaluate the performance of ASV systems. The authors notice a significant decline in performance for underrepresented groups in training (females and speakers with nonnative English accents). They show that a simple data balancing approach mitigates this bias on minority groups without sacrificing performance on the majority groups.

In-processing methods are based on integrating fairness into the model during training by introducing fairness constraints. Shen et al. [177] demonstrate that imbalanced gender representation in training sets can result in model unfairness. To address this issue, the authors propose training group-adapted encoders to extract gender-specific embeddings. This technique not only improves fairness but also enhances the overall system utility. Fairness evaluations are based on the differences in EER between the genders.

Jin et al. [178] suggest an adversarial reweighting training technique to reduce bias in ASV systems. By employing an adversarial network, this approach automatically identifies underperforming groups and adjusts their impact on the training loss. The proposed method improves both performance discrepancy across gender and nationality demographic groups and overall performance of the ASV system without necessitating explicit information about group membership during the training.

Peri et al. [175] propose a solution to address gender biases in ASV systems by using adversarial and multi-task learning techniques. By combining these technique together, the method aims to generate demographic-aware speaker embeddings to reduce bias. The authors use fairness discrepancy rate (FDR) [163] metric that weights absolute discrepancy in FAR and FRR between demographic groups.

Hutiri et al. [179] highlight the importance of mitigating bias with inclusive evaluation datasets and developed design guidelines for these datasets. The authors recommend the use of representative datasets that mirrors the diversity of the demographics of the population. This representation should be maintained at both the speaker and utterance levels to ensure fair speaker verification assess-

ments evaluation across demographic groups.

2.3 Conclusion

In this comprehensive chapter, we have presented privacy enhancing technologies, focusing on secure multi-party computation, homomorphic encryption, and differential privacy mechanisms. We have examined the state-of-the-art use of PETs in biometrics, highlighting their role in preserving privacy to be compliant with the GDPR. We have further presented the challenges associated with integrating these techniques with machine learning-based speaker verification systems.

Additionally, we have introduced disentanglement representation learning as an emerging privacy preservation technique within machine learning. Furthermore, we have presented state-of-the-art applications of these techniques in biometrics, showcasing their efficacy in safeguarding sensitive data while maintaining utility.

Moreover, we have explored the concept of fairness in biometrics and discuss fairness assessment measures. We have then presented bias assessment and mitigation techniques tailored specifically for speaker verification systems.

Overall, this thorough review of existing literature aims to foster a comprehensive understanding of both well-established and emerging technologies, promoting data privacy, ethical use, and GDPR compliance to data privacy and fairness principles.

Chapter 3

Privacy-Preserving Voice Anti-Spoofing based on Secure Multi-Party Computation

In this chapter, we introduce the security vulnerabilities of ASV systems and the countermeasures to the spoofing threats. We further present the challenges faced by cloud-based anti-spoofing systems in secure speaker authentication, particularly when balancing security with privacy. We then introduce PRIVASP, the first proposed solution for privacy-preserving voice biometric anti-spoofing. PRIVASP is based on adapting an MPC technique to a shallow neural network anti-spoofing system.

3.1 Automatic Speaker Verification System Security Issues and Countermeasures

Despite the reinforced security and user convenience ASV systems provide, just like other biometric systems, they are susceptible to a range of attacks, categorized following the ISO/IEC 30107-1 standards [180, 181]. Figure 3.1 details these threats which vary from sensor-level threats, such as using a microphone to capture and replay a voice of a legitimate user, to more sophisticated methods that can compromise the authentication decision. The most critical vulnerabilities in ASV systems are physical access (PA) attacks (at the microphone level) and logical access (LA) attacks (at the acquisition stage prior to signal processing). PA and LA attacks, also referred to as spoofing or presentation attacks, are direct

attacks and do not require access to the core system [182]. These types of attacks are easier to perform compared to attacks at other system levels, thereby posing a greater threat.

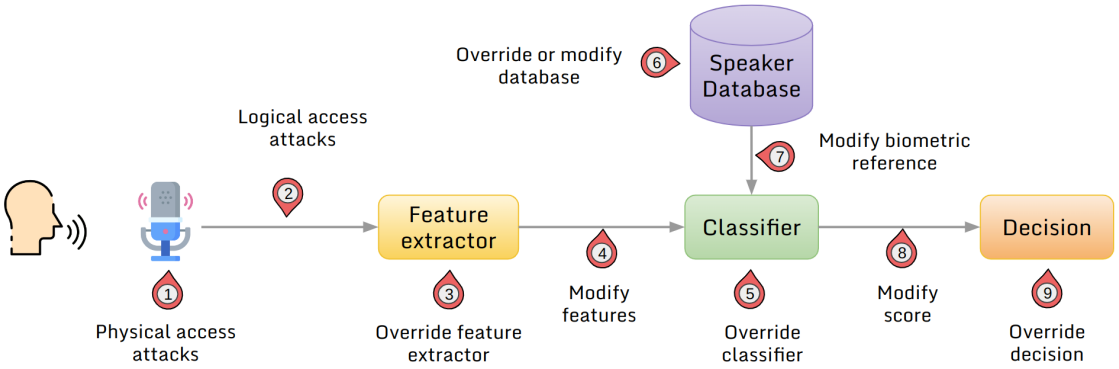


Figure 3.1: List of attacks at different vulnerable points on an ASV system

Spoofing attack poses a great security threat and need to be urgently addressed. Researchers have focused efforts on identifying ASV systems vulnerabilities and developing adequate countermeasures (CMs), also known as presentation attack detection (PAD) solutions in the ISO/IEC 30107-1 standard [180]. The four biennial ASVspoof challenges [183–186] have promoted the development and implementation of CMs to automatically detect and counter ASV system spoofing attacks. These challenges shared comprehensive databases containing both bona fide (i.e. genuine) and spoofed utterances, baseline systems, as well as ASV and CMs evaluation protocols. Anti-spoofing measures are incorporated [185], as depicted in Figure3.2. The spoofing countermeasure system produces scores to be combined with the scores of the ASV system in order to ensure the authenticity of the speech and confirm the identity of the speaker prior to allowing access.

3.2 Privacy Threats of Cloud-Based Spoofing Countermeasures Systems

Cloud-based anti-spoofing systems serve as a decentralized solution to ensure secure speaker verification, thwarting fraudsters who impersonate legitimate enrolled subjects to unlawfully access resources secured by speaker verification systems. These cloud-based services necessitate the transmission of speech data, which may traverse potentially vulnerable networks. Besides, cloud-based servers are vulner-

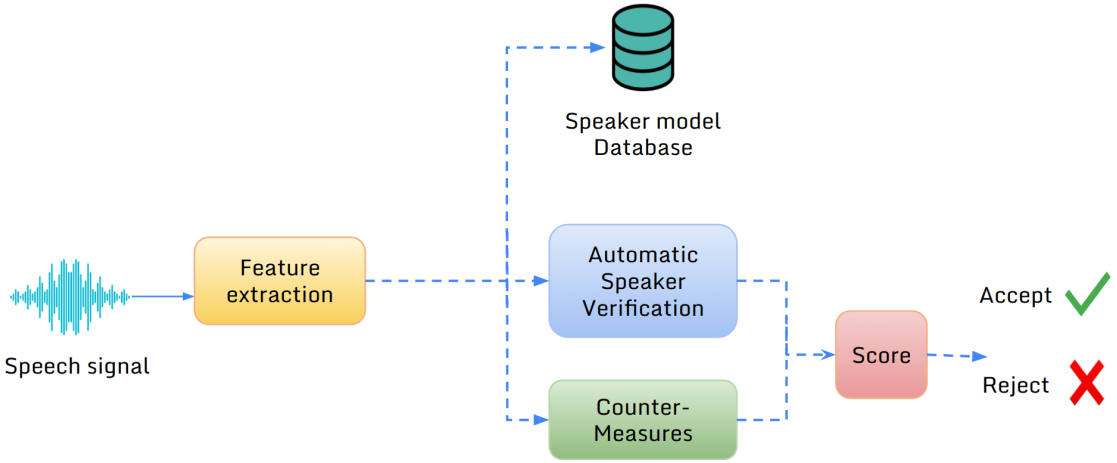


Figure 3.2: Representation of the Automated Speaker Verification (ASV) and Countermeasure (CM) Systems.

able to data breach. In January 2018, a significant security breach was reported in Aadhaar, the world's largest ID database, compromising the personal and biometric information of over 1.1 billion Indian citizens [1].

Consequently, privacy legislation globally, including the GDPR in Europe, now requires the implementation of protective measures. It is imperative to consider alternative mechanisms that prioritize user privacy without compromising security. Local processing of speech data presents itself as a potential solution. By confining data processing to the user's device, the risks associated with data transmission across networks and cloud-based server vulnerabilities can be significantly mitigated. This method, however, introduces the necessity of deploying service provider-developed models onto users' personal devices. These models, being the product of extensive data aggregation and intensive research, represent a substantial intellectual and developmental investment. The proprietary nature of these models raises legitimate concerns for service providers about the potential exposure of their intellectual property. The threat of intellectual property (IP) theft, whether through reverse engineering or other means, makes service providers understandably cautious about distributing their models for local processing. Despite the clear privacy benefits that local processing affords to users, the practicality of this solution from the standpoint of the service providers remains contentious.

[1] `http://tinyurl.com/bdfefpnx`

The balance between ensuring user privacy and protecting the IP of service providers is delicate and complex. In this context, the deployment of privacy-enhancing technologies becomes pivotal.

3.3 Proposed system: PRIVASP

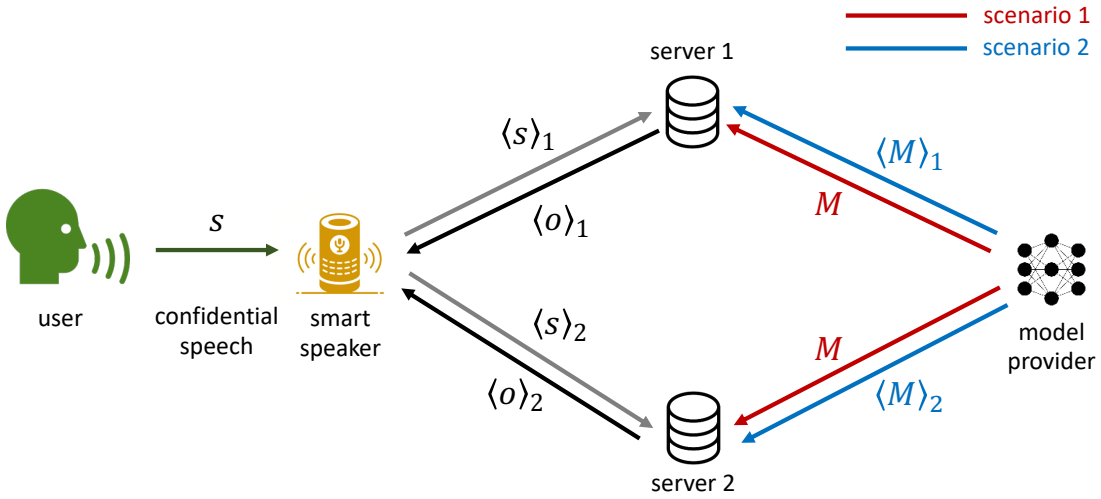


Figure 3.3: Scenario 1: PRIVASP with Model privacy against the client (red arrows). Scenario 2: PRIVASP with model privacy against both the client and the cloud servers (blue arrows).

In our research, we focus on a scenario where a user attempts to authenticate with an ASV system, like Google Home, using his/her voice. A critical aspect of this process is the ability of the ASV system to determine if the voice input is authentic or a spoofing attempt. To assist in this, the ASV system employs an external anti-spoofing service equipped with a specialized NN model, hosted on cloud servers.

The challenge arises from the need to protect sensitive information: the user's voice data must be kept secret from the anti-spoofing service and the cloud servers, while the service provider also seeks to safeguard its anti-spoofing model, considering it a valuable asset. Furthermore, the cloud servers themselves could pose a threat to the confidentiality of the model.

To tackle these privacy issues, we propose the use of a secure 2PC technique based on additive secret sharing, where all parties involved are assumed to be hon-

est but curious (refer to Chapter 2.1.1.1). As previously discussed in Chapter 2.1.3, implementing a secure 2PC scheme in existing ASV systems is not straightforward due to their complexity, especially those using deep neural networks with many non-linear operations, as in [9, 187]. These complexities could lead to prolonged processing times under MPC. Also, the limitation of MPC to integer calculations complicates matters further, as these ASV systems usually process real numbers, and converting these to integers can reduce accuracy. Additionally, executing non-linear operations like comparisons in MPC requires specific approximations.

Therefore, our solution is to design a new spoofing countermeasure system compatible with MPC. We introduce PRIVASP, a system with a shallow NN architecture that not only facilitates the integration of MPC but also maintains spoofing detection accuracy. We detail the performance and efficiency of PRIVASP in Section 3.4.4.

In the following two sections we explain how MPC is used in PRIVASP and suggest two versions of it as illustrated in Figure 3.3. In both versions, the speech is secretly shared among the two cloud servers. On the other hand, in the first version, the anti-spoofing service fully trusts the cloud and sends the model parameters in clear. In the second version, in addition to the input, the model parameters are also secretly shared. The first version is represented in red in Figure 3.3, where the same model is shared between the servers, and the second version is represented in blue in which the model is secretly shared.

3.3.1 PRIVASP with Model privacy against the client

In our proposed system, the process begins with the client, who has a voice matrix representing the user's voice input. To maintain confidentiality, the client splits this voice matrix into two separate parts using additive secret sharing (Section 2.1.1.1.2). This method involves creating two distinct secret shares, each of which is sent to a different server. These two servers are set up to be non-colluding cloud servers.

Upon receiving their respective shares, each server conducts a spoofing detection task using the same anti-spoofing model, denoted as M. This model is trained to identify whether the voice input is genuine or a spoofing attempt. A key aspect of this setup is that the anti-spoofing model, which is a crucial asset in this scenario, is kept hidden from the client. The client only provides the input data

but does not have access to or knowledge of the model itself. The parameters of the model remain exclusively within the servers, ensuring that the client cannot reverse-engineer or otherwise obtain insights about the model.

However, an important consideration arises in this setup. In this initial design of PRIVASP, the model, while protected from the client, is fully accessible to the cloud servers which perform the classification task. This situation poses a risk in terms of IP: if the model is a proprietary asset, exposing it to the cloud servers could lead to unauthorized access or duplication.

Recognizing this challenge, we propose an advanced version of PRIVASP in the subsequent section of our work. This second version of the system aims to fortify the confidentiality of the anti-spoofing model and extending privacy protection to encompass not only the client but also the cloud servers. This enhanced approach seeks to ensure that the parameters of the model are securely shielded from all external parties, thereby better safeguarding the IP and maintaining the integrity of the anti-spoofing service.

3.3.2 PRIVASP with model privacy against both the client and the cloud servers

In this refined version of our system, we carefully manage the privacy of both the user's voice data and the pre-trained anti-spoofing model using a secure 2PC protocol. Initially, the parameters of the pre-trained model are split into secret shares and securely distributed to two non-colluding servers, ensuring no single server has access to the complete model. Simultaneously, the client's voice data is also divided into secret shares and sent to these servers. As part of the secure 2PC protocol, the servers, each possessing a piece of the model and the data, collaboratively compute the spoofing detection task. This computation is performed in a way that the servers themselves cannot access or reconstruct the full model or the original data. The integrity of the model is thus maintained, and the IP is protected. The final step involves sending the result of the spoofing detection back to the smart device. While the servers perform the computation, they remain oblivious to both the input data's nature and the computation's outcome. This end-to-end process ensures that the client's data and the model's confidentiality are preserved, while still providing accurate and reliable spoofing detection.

3.4 Experimental setup

This section outlines the ASVspoof 2019 LA database, evaluation metrics, baselines of the challenge, competing state-of-the-art countermeasures, and PRIVASP implementation details. It is important to note that the countermeasure systems examined in the experiments are individual systems. This means that they are not a fusion of multiple individual systems, but are standalone in their operation and analysis.

3.4.1 ASVspoof 2019 LA database

Experiments were carried out using the publicly available ASVspoof 2019 LA database [2], a component of the ASVspoof 2019 challenge delineated in subsection 3.1.

Subset	Number of speakers		Number of utterances	
	Male	Female	Bona fide	Spoofed
Training	8	12	2580	22800
Development	4	6	2548	22296
Evaluation	21	27	7355	63882

Table 3.1: Statistics of the database used in ASVspoof 2019 challenge Logical Access partition.

The ASVspoof 2019 LA subset, presented in Table 3.1, is divided into three disjoint partitions: training, development, and evaluation. Each partition includes bona fide utterances and spoofed utterances generated using a variety of advanced voice conversion (VC), text-to-speech (TTS), and hybrid (VC-TTS) algorithms, with a total of nineteen algorithms. TTS algorithms generate spoken output from text input, whereas VC algorithms transform input source speech to to resemble that of a target speaker. The training and development sets include four TTS algorithms and two VC algorithms. The evaluation set includes seven TTS algorithms, three VC algorithms, and three hybrid algorithms. Further details on each attack algorithm can be found in [185]. In the 2019 LA evaluation set, two attacks are categorized as known, as they employ identical algorithms to those used in the training and development sets, with distinct utterances and speakers. The

[2] `https://datashare.ed.ac.uk/handle/10283/3336`

remaining eleven attacks in the evaluation set are classified as unknown. Although these unknown attacks may employ similar techniques to the known ones, their full algorithms differ.

3.4.2 Evaluation metrics

Spoofing detection in ASV systems is formulated as a binary classification task, with classifiers assigning scores to input trials. Trials with scores exceeding a pre-fixed threshold, τ_{cm}, are deemed genuine; others are classified as spoofed (i.e. imposter). Two speaker verification evaluation metrics where proposed in this challenge.

3.4.2.1 Equal Error Rate

Evaluation employs the Equal Error Rate (EER), where the false acceptance rate (P_{fa}^{cm}) and the false rejection rate (P_{miss}^{cm}) converge. Defined as follows, these rates are computed using the Bosaris toolkit:

$$P_{fa}^{cm}(\tau_{cm}) = \frac{\#\text{spoofed trials with CM scores} > \tau_{cm}}{\#\text{total spoofed trials}} \tag{3.1}$$

$$P_{miss}^{cm}(\tau_{cm}) = \frac{\#\text{bona fide trials with CM scores} \leq \tau_{cm}}{\#\text{total bona fide trials}} \tag{3.2}$$

Here, False accepts (FAs) occur when spoofed trials are mistakenly accepted, and False rejects (FRs) when genuine trials are mistakenly rejected. Despite its recent deprecation in ISO/IEC standards [180, 181], EER remains a prevalent metric within the speaker recognition community due to its intuitive interpretation. However, as a parameter-free metric, devoid of priors or detection costs, EER does not fully reflect practical performance scenarios.

3.4.2.2 Tandem Detection Cost Function

Recognizing the need for a more representative metric, the minimum tandem Detection Cost Function (min t-DCF) was introduced in 2018 by Kinnunen et al. [188, 189]. This metric provides a more holistic measure of the impact of spoofing and CMs on the reliability of an ASV system. This metric accounts for the interdependent operations of the ASV and CM within the same framework. Hence, min t-DCF is favored for its ability to capture the practical performance of ASV

systems under potential spoofing conditions. The min t-DCF is defined as:

$$\text{min t-DCF} = \min_{\tau} \frac{C_0 + C_1 P_{miss}^{cm}(\tau_{cm}) + C_2 P_{fa}^{cm}(\tau_{cm})}{C_0 + \min(C_1, C_2)} \tag{3.3}$$

where C_0, C_1, and C_2 are hyper-parameters derived from ASV scores and the priors of target (i.e. mated), non-target (i.e. non-mated), and spoofed trials, in addition to ASV and CM detection costs:

$$C_0 = \pi_{tar} C_{miss}^{asv} P_{miss}^{asv} + \pi_{non} C_{fa}^{asv} P_{fa}^{asv} \tag{3.4}$$

$$C_1 = \pi_{tar} C_{miss}^{cm} - (\pi_{tar} C_{miss}^{asv} P_{miss}^{asv} + \pi_{non} C_{fa}^{asv} P_{fa}^{asv}) \tag{3.5}$$

$$C_2 = \pi_{spoof} C_{fa}^{cm} P_{fa}^{spoof} \tag{3.6}$$

The min t-DCF metric quantifies the combined security effectiveness of ASV and CM systems, with values ideally ranging between 0, denoting perfect system performance, and 1, indicating reduced protection performance against spoofing attacks. The cost parameters C_{miss}^{asv}, C_{fa}^{asv}, and C_{miss}^{cm} correspond to the costs of target trial rejections, non-target acceptances, and bona fide trial rejections by the ASV and CM systems, respectively. For an in-depth discourse on the computation and implications of these parameters, readers are directed to [188, 189].

3.4.3 ASVspoof 2019 baselines and post-evaluation systems

In the ASVspoof 2019 Challenge [185], participants were provided with two baseline countermeasure (CM) systems. Baseline B01 employs constant Q cepstral coefficients (CQCCs) [190], which are a set of features derived from a logarithmic frequency scale where the number of frequency bins per octave is fixed at 96, enhancing resolution in lower frequencies. The resampling period is specified as 16, which determines the rate at which the frequency spectrum is sampled. The feature vector comprises 29 static coefficients, including the zeroth coefficient, capturing the spectral envelope of the signal. These static features are further enhanced by their first and second temporal derivatives, known as delta and delta-delta coefficients, providing dynamic information about the trajectory of the cepstral features

over time. Consequently, this results in a comprehensive 90-dimensional feature vector, utilized in conjunction with a 512-component Gaussian Mixture Model (GMM) serving as the back-end classifier [191, 192].

Baseline B02 utilizes linear frequency cepstral coefficients (LFCCs), which linearly sample the frequency spectrum between 30 Hz and 8 kHz, using a 512-point Discrete Fourier Transform (DFT) applied to signal frames of 20 ms with a 50% overlap between consecutive frames. This approach captures both the stationary and transitional properties of the speech signal. Similar to B01, LFCCs include 19 static coefficients and the zeroth coefficient, with the addition of delta and delta-delta coefficients to encapsulate the temporal dynamics, resulting in a 60-dimensional feature vector. This feature set is also analyzed by a 512-component GMM classifier [193].

Both baseline systems are trained on bona fide and spoofed utterances from the ASVspoof 2019 training dataset using an expectation-maximization (EM) algorithm. The model outputs are log-likelihood ratios that differentiate between genuine and spoofed speech. The corresponding Matlab package for both baselines can be obtained from the ASVspoof website [3].

In addition to these baselines, three advanced systems were evaluated post-challenge for comparison. These include the high-spectral resolution LFCC system with a traditional GMM classifier (LFCC-GMM) [194], RawNet2 [187], and ResNet18-SP [9] systems. Notably, the latter two are sophisticated DNN models, comprising millions of parameters, indicative of their complexity and depth in feature representation.

3.4.4 Implementation details of PRIVASP

PRIVASP is designed for collaborative spoofing detection, engaging a client (such as a home assistant) and a CM model provider, alongside two non-colluding servers (e.g. Alice and Bob) to ensure secure and private computation. The client's input (and the model's weights in the second scenario) is secretly shared between Alice and Bob, while a third party, the crypto Provider, generates random numbers for secure multiplication, without owning any shares or colluding with the other parties.

PRIVASP uses LFCCs as the front-end feature extractor. Audio signals are

[3] `https://www.asvspoof.org/index2019.html`

processed by segmenting the speech waveform into overlapping frames, precisely using a 30 ms window with a 15 ms shift. From each frame, 30 LFCCs are derived, focusing on the first 1500 ms of each utterance; for utterances less than 1500 ms, repetition is employed to meet the length requirement. The resulting feature matrix is then vectorized into a column vector comprising 2970 elements.

In the back-end, a shallow NN with a single hidden layer featuring the rectified linear unit (ReLU) activation function is employed. We evaluate two variants of PRIVASP: PRIVASP-1024 and PRIVASP-512—distinguished by the number of neurons in their hidden layers, 1024 and 512, respectively. Figure 3.4 illustrates the architecture of the shallow NN used in PRIVASP-1024.

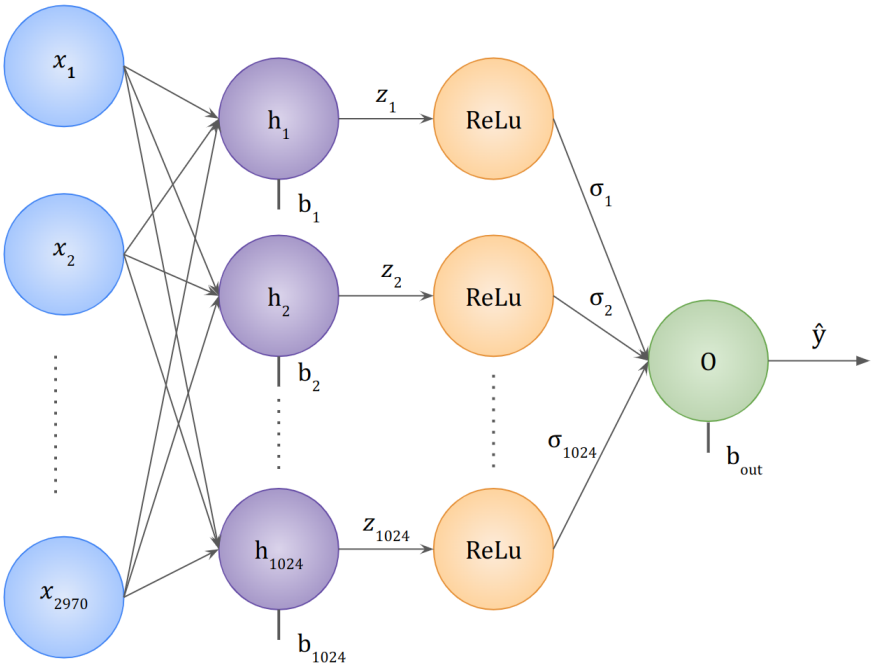


Figure 3.4: Architecture of the PRIVASP-1024 shallow neural network.

The PRIVASP-1024 NN processes input features $x_1, x_2, \ldots, x_{2970}$ through a single hidden layer to produce the output $\hat{y}$. In a fully connected layer, each neuron is interconnected with all neurons in the previous layer, allowing the network to learn complex patterns from the input data. Each hidden neuron computes a weighted sum of the inputs, yielding $z_i = \sum_j w_{ij}x_j + b_i$, where w_{ij} are the weights, x_j are the input features, and b_i is the bias. The z_i values are then passed through a ReLU, defined as $\text{ReLU}(z_i) = \max(0, z_i)$, introducing non-linearity and

enabling the identification of patterns by activating neurons when input exceeds a threshold, which helps the network in recognizing diverse features. The training of the network levarages the adaptive moment estimation (Adam) optimizer to fine-tune the weights, aiming to minimize the binary cross-entropy loss function. This optimizer dynamically adjusts learning rates, using estimates of lower-order moments to enhance convergence. The binary cross-entropy loss function, essential for binary classification tasks, quantifies the discrepancy between the predicted outputs and actual labels and guides the model towards more accurate predictions. Model selection favors the iteration exhibiting the lowest loss value during training. The training of the spoofing detection model is performed on the model provider's side in clear, without applying MPC protection.

The experimental setup consisted of a PC equipped with an Intel i5-9400F 6-core processor at 2.9 GHz, an NVIDIA GeForce GTX 1050 GPU with 4GB of memory, and 64GB of RAM. PyTorch framework[4] was employed for NN construction. Pysyft library[5], which integrates MPC capabilities within PyTorch, was used for PRIVASP implementation after the training and during the employment of the CM [195, 196].

In the secure 2PC context within PySyft, operations necessary for neural computation, such as addition, multiplication, and the non-linear ReLU, are securely performed. PySyft's MPC framework employs protocols like SPDZ [197, 198] and SecureNN [86] to execute these computations while ensuring data privacy. To comply with the MPC's requirements, real numbers in the input layer are translated into integer-based fixed-precision numbers using PySyft's *FixedPrecisionTensor*, which encodes floating-point numbers and maintains the radix point location. For example, the floating-point number 0.123 with precision 2 is rounded to the integer 12.

The ReLU function is securely computed through the SecureNN protocol. Alice and Bob engage in a series of sub-protocols that allow them to compare shared values and compute the most significant bit (MSB) without revealing the actual input value. If the MSB indicates a negative input, ReLU outputs zero; for non-negative inputs, it outputs the value itself. This process is conducted without disclosing individual shares, thereby preserving data privacy and enabling secure

[4] `https://pytorch.org/`

[5] `https://github.com/OpenMined/PySyft`

neural network operations. For in-depth details on these sub-protocols, readers may consult [86].

The output $\hat{y}$, illustrated in Figure 3.4, is mathematically represented by:

$$\hat{y} = \sum_{i=1}^{1024} w_{oi} \cdot \sigma_i + b_{\text{out}},$$

where w_{oi} is the weight from the i-th hidden neuron to the output neuron, σ_i denotes the output of the ReLU activation function $\text{ReLU}(z_i)$, and b_{out} is the output neuron's bias.

In a secure 2PC setting, the output is held as separate shares by Alice and Bob, noted as $\hat{y}_A$ and $\hat{y}_B$. The client combines these shares to form the final score:

$$\text{Score} = \hat{y}_A + \hat{y}_B.$$

This computed score, when compared with a predefined threshold, classifies the speech as genuine if it exceeds the threshold or as spoofed otherwise.

3.5 Results

The evaluation follows a threefold objective: i) analyzing the performance of countermeasures, ii) assess privacy-preserving algorithms, and iii) evaluate the computational costs.

In Tables 3.2 and 3.3, we present the experimental results for the baseline systems B01 and B02, alongside advanced post-evaluation models namely the high-spectral-resolution LFCC, RawNet2, and ResNet18-SP, previously presented in section 3.4.3, as well as our proposed PRIVASP-1024 and PRIVASP-512 systems. The tables detail performance metrics in terms of pooled EER and min t-DCF.

The term *plaintext* in these tables refers to the conventional setting where no privacy preservation techniques are applied, and all computations are performed in the clear. This serves as a benchmark to evaluate the spoofing detection performance of the models without the added complexity of secure two-party computation. In this context, *PRIVASP plaintext* indicates the performance of our shallow neural network when operating without applying the privacy-preserving and without distributing data among the two non-colluding servers.

To rigorously assess the spoofing detection performance under privacy con-

system	type	EER [%]	min-tDCF
B01 [190]	plaintext	0.43	0.0123
B02 [193]	plaintext	2.71	0.0663
LFCC-GMM [194]	plaintext	0.00	0.0000
RawNet2 [187]	plaintext	1.09	0.0362
ResNet18-SP [9]	plaintext	0.07	0.0018
PRIVASP-1024	plaintext	0.00	0.0000
	scenario 1	0.00	0.0000
	scenario 2	0.00	0.0000
PRIVASP-512	plaintext	0.00	0.0000
	scenario 1	0.00	0.0000
	scenario 2	0.00	0.0000

Table 3.2: Performance for the ASVspoof 2019 LA development partition in terms of pooled EER and min t-DCF for the two baselines, B01 and B02, the high-spectral-resolution LFCC, RawNet2, ResNet18-SP and our proposed PRIVASP-1024 and PRIVASP-512 systems. PRIVASP systems are also evaluated in privacy-preserving scenario 1 and 2.

straints, we conducted separate experiments for PRIVASP systems in the two scenarios, previously explained in section 3.3. As shown in Table 3.2, associated with the development partition, both PRIVASP-1024 and PRIVASP-512 exhibit exemplary performance, achieving perfect scores even under the stringent conditions of privacy-preserved scenarios. In the evaluation partition, represented in Table 3.3, PRIVASP-1024 marginally outperforms PRIVASP-512 in plaintext conditions with an EER of 7.03% and a min-tDCF of 0.1485. Performance remains consistent across the two privacy-preserved scenarios, indicating that applying 2PC did not compromise the accuracy of the system. PRIVASP systems demonstrate superior performance against the baselines B01 and B02, as well as the RawNet2 system as shown in Table 3.3.

Efficiency is a crucial factor in the practical deployment of spoofing CM systems. As indicated in Table 3.4, we assess the average inference time, measured in milliseconds, to determine whether an utterance is bona fide or spoofed. Our analysis reveals that PRIVASP-512 consistently outpaces PRIVASP-1024 in terms of inference time, due to its streamlined architecture with fewer neurons. In ciphertext scenario 2, PRIVASP-512 achieves an impressive detection time of approximately

system	type	EER [%]	min-tDCF
B01 [190]	plaintext	9.57	0.2366
B02 [193]	plaintext	8.09	0.2116
LFCC-GMM [194]	plaintext	3.50	0.0904
RawNet2 [187]	plaintext	5.54	0.1547
ResNet18-SP [9]	plaintext	6.82	0.1140
PRIVASP-1024	plaintext	7.03	0.1485
	scenario 1	7.02	0.1481
	scenario 2	7.02	0.1481
PRIVASP-512	plaintext	7.10	0.1549
	scenario 1	7.13	0.1550
	scenario 2	7.13	0.1550

Table 3.3: Performance for the ASVspoof 2019 LA evaluation partition in terms of pooled EER and min t-DCF for the two baselines, B01 and B02, the high-spectral-resolution LFCC, RawNet2, ResNet18-SP and our proposed PRIVASP-1024 and PRIVASP-512 systems. PRIVASP systems are also evaluated in privacy-preserving scenario 1 and 2.

208ms per utterance, compared to around 350ms for PRIVASP-1024. These times are competitive with the baseline B01 system when operating in plaintext, and they are well within the bounds acceptable for real-time application requirements.

In the less stringent ciphertext scenario 1, where model privacy is not a concern, the PRIVASP systems demonstrate even greater efficiency. PRIVASP-1024 and PRIVASP-512 report detection times of roughly 95ms and 60ms, respectively, surpassing the plaintext performance of the baseline systems B01, B02, and LFCC-GMM. The PRIVASP variants in plaintext mode exhibit comparable efficiency to the more complex deep learning models RawNet2 and ResNet18-SP.

It is important to note that these efficiency metrics for PRIVASP are obtained without the aid of GPU acceleration, which is a standard practice for deep learning models. This highlights the optimized nature of PRIVASP's neural network design, tailored for efficient computation while still ensuring robust privacy-preserving capabilities in a multi-party computation context.

system / type	**PRIVASP-1024**	**PRIVASP-512**	B01 [190]	B02 [193]	LFCC-GMM [194]	RawNet2 [187]	ResNet18-SP [9]
plaintext	2.8	2.7	339.9	89.9	100.6	12.0	2.8
scenario 1	95.8	59.9	-	-	-	-	-
scenario 2	349.6	208.1	-	-	-	-	-

Table 3.4: Average inference time in ms per utterance.

3.6 Summary

This chapter introduces PRIVASP, the first privacy-preserving solution to voice anti-spoofing. Instead of treating privacy as an afterthought, PRIVASP is fundamentally built with a *privacy by design* approach. This pioneering approach ensures that the spoofing countermeasure is inherently tailored to align with the capabilities of 2PC. As a result, PRIVASP not only upholds rigorous standards of privacy protection but also guarantees efficient spoofing detection. The system is underpinned by a carefully architected shallow neural network, equipped with a single layer and the ReLU activation function, making it adept for MPC environments.

Our experiments were performed on the ASVspoof 2019 Logical Access database, where we explored two distinct operational scenarios. These scenarios were differentiated based on whether the spoofing countermeasure service providers opted to disclose their models to the cloud service provider or not.

The outcomes of our experiments underscore the remarkable efficiency of PRIVASP, even within the traditionally computation-intensive realm of MPC. By innovatively designing a shallow neural network, PRIVASP aligns seamlessly with our secure 2PC framework, effectively overcoming the usual computational overhead associated with such techniques. The result is a system that not only excels in real-time spoofing detection but also operates with remarkable efficiency. This breakthrough represents a significant step in simultaneously achieving the dual goals of robust user privacy protection and safeguarding the service provider's intellectual property by keeping the model parameters private.

Chapter 4

Protecting Gender in Voice Biometrics Based on Differential Privacy and Adversarial Training

This chapter presents a novel privacy-preserving speaker verification solution focusing on concealing gender attributes from speaker embeddings while retaining identity of speakers. This approach combines two techniques: adversarial training and differential privacy previously described in Chapter 2 (Section 2.1.4.1 and Section 2.1.2.2, respectively).The chapter is structured as follows. First, the motivation behind the development of this privacy-preserving solution is discussed. Next, the methods employed for gender concealment are presented. This includes the implementation of the gender adversarial auto-encoder (Gender-AAE) and the Gender-AAE with Laplace noise layer. Following this, the experimental evaluation and results are presented. This encompasses details regarding the databases used, the experimental settings employed, and the analysis of gender-neutral speaker representations. Finally, a summary of the key findings and contributions of this chapter is provided.

4.1 Motivation

The digitization of identity verification through voice biometrics emphasizes the need for privacy-preserving technologies. As discussed in Chapter 1.3, the human voice inherently encapsulates numerous physiological and psychological traits, including gender, a particularly sensitive attribute. The indiscriminate exposure

of such traits can lead to significant privacy risks. Hence, it is imperative to implement robust protection measures against potential misuse or unauthorized analysis.

Consider the scenario depicted in Figure 4.1, where an online education platform employs speaker verification for user authentication. The platform adheres to privacy and data minimization principles by maintaining gender-neutral profiles and using voice prints solely for identity verification. This approach ensures equal treatment of users and prevents content personalization based on gender, aligning with a commitment to privacy requirement by the GDPR. However, this setup is not devoid of risks. If a staff member responsible for maintaining the speaker verification system decides to analyze the voice data using a pre-trained gender classifier, they might uncover the user's gender. Such unauthorized gender inference could inadvertently bias the educational content, contravening GDPR laws that advocate for minimal data usage and fostering non-discriminatory practices.

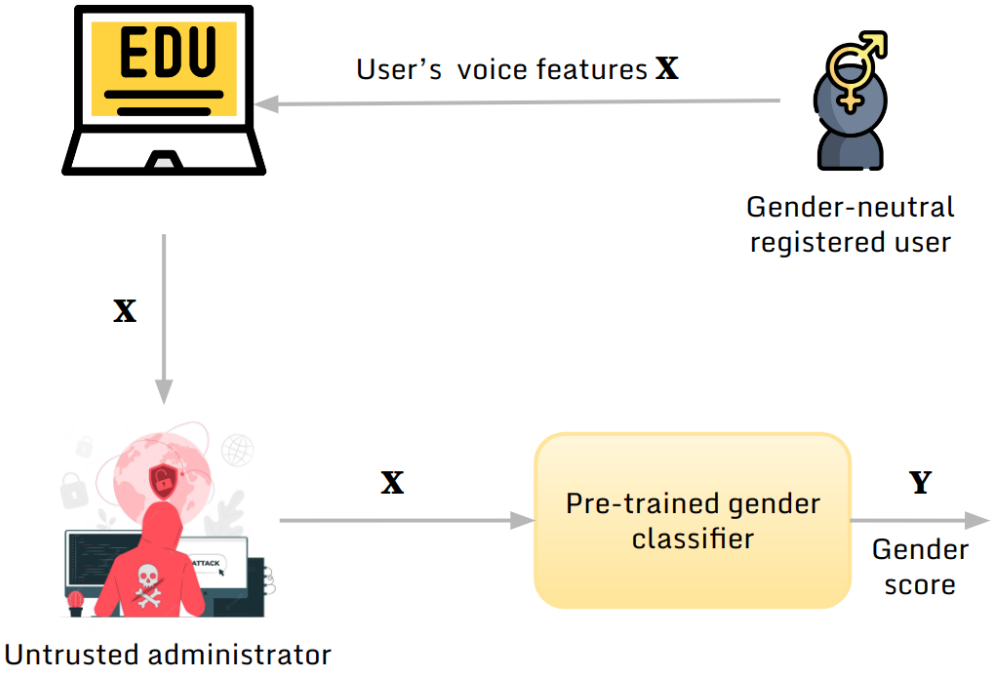


Figure 4.1: The risk of gender inference by an untrusted e-learning website administrator.

Traditional cryptographic methods such as MPC present a clear trade-off: they secure data but often at the expense of compatibility with existing ML models.

These models must often be completely redesigned to operate efficiently within the constraints of MPC, as presented in chapter 3. Additionally, not all operations are supported by MPC and HE, limiting their applicability. MPC, while robust, is known for its computational complexity, and HE is marked by its time-intensive computations, as discussed in Chapter 2.1.3. There is also a complex knowledge required to effectively implement these solution that are not straightforward. Consequently, there is a high demand for more accessible solutions within the biometric community. These solutions should offer lower computational and communicational complexities and support a wider range of operations to facilitate their adoption and integration into existing systems.

Data obfuscation techniques offer a refined alternative, striking a balance that protects sensitive information without compromising the utility of the data. This balance is critically important in the field of speaker verification, where verifying a person's identity must be judiciously weighed against the crucial need to protect sensitive attributes like gender.

To tackle this challenge, this chapter introduces the first initiative that effectively utilizes differential privacy, a concept traditionally associated with data anonymization (see Chapter 2.1.2), to protect gender information in speaker verification, all while preserving individual identity. Our innovative approach leverages the strengths of DP in conjunction with an adversarial auto-encoder framework. This distinctive integration allows us to uphold the integrity of individual identity features within speaker embeddings. In this work, DP is employed not to obscure but to safeguard these identity traits, specifically focusing on the concealment of gender traits.

4.2 Gender concealment

In this section, we introduce the building blocks of the suggested technique for obscuring gender information. Initially, we examine the structure of the AAE, which largely aligns with the system proposed in [139], and discuss its constraints regarding the task of gender concealment. Subsequently, we demonstrate the integration of the AAE with LDP to enhance the efficacy of the model in minimizing gender information within speaker embeddings. This combination offers a configurable balance between privacy and utility, backed by solid theoretical assurances presented in Chapter 2.1.2.2.

4.2.1 Gender-Adversarial Auto-Encoder

Let $\mathbf{x}$ be an embedding representing a speaker identity. The primary objective of a gender adversarial auto-encoder is to refine $\mathbf{x}$ into a new representation, $\tilde{\mathbf{x}}$, which preserves the speaker's identity while stripping away gender characteristics. This transformation is achieved through a series of feed-forward neural network modules, each with a specialized function. The first step of the process involves compressing $\mathbf{x}$ into a latent representation $\mathbf{z}$ using an encoder function $e_{\phi_1}(\mathbf{x})$. This encoder module, parameterized by ϕ_1, reduces the dimensionality from the original space $\mathbb{R}^d$ to a smaller latent space $\mathbb{R}^l$. In this phase, the encoder's primary task is to maintain the essential characteristics of the speaker's identity in $\mathbf{z}$ while concealing gender-specific features.

The next key component in the Gender-AAE system is the adversarial discriminator module, denoted as $a_\theta(\cdot)$. The discriminator initially undergoes training to identify the gender of the speaker from the latent representation $\mathbf{z}$. This training phase is crucial as it defines the encoder's target level of gender obfuscation. The optimization of the discriminator's parameters, θ, focuses on minimizing the following discriminative loss function:

$$\mathcal{L}_{disc}(\mathbf{x}, y, \theta \mid \phi_1) = -y \log(a_\theta(\mathbf{z})) - (1-y) \log(1 - a_\theta(\mathbf{z})) \tag{4.1}$$

where $y \in \{0, 1\}$ is the binary gender label, with 0 representing male and 1 representing female. The term $a_\theta(\mathbf{z})$ represents the probability assigned by the discriminator to the likelihood of $\mathbf{z}$ being generated by a female speaker.

The interaction between the encoder and the discriminator can be likened to a strategic two-player game. The encoder is focused on generating a latent representation $\mathbf{z}$ that effectively masks gender details, making it challenging for the discriminator to accurately predict the speaker's gender. Conversely, the discriminator analyzes the compact data $\mathbf{z}$ with the aim of uncovering any concealed gender information. If the discriminator succeeds in detecting gender, it indicates a need for the encoder to further refine its gender-hiding techniques. In practice, gender-related information is concealed by training the encoder to 'fool' the discriminator, with both networks optimizing the same objective as (4.1) but with the distinction that the predicted probability by the discriminator is inverted:

$$\mathcal{L}_{adv}\left(\mathbf{x}, y, \phi_1 \mid \theta\right) = -y \log\left(1 - a_\theta\left(\mathbf{z}\right)\right) - \left(1 - y\right) \log\left(a_\theta\left(\mathbf{z}\right)\right) \tag{4.2}$$

Finally, a decoder module $d_{\phi_2}(\cdot)$ attempts to reconstruct the original input embedding from $\mathbf{z}$. The role of the decoder is to guarantee that the reconstructed embedding can still be used for other tasks (i.e. speaker verification in this case) despite the suppression of gender-related attributes. Thus, the auto-encoder is optimized end-to-end according to a further reconstruction objective: the cosine distance between the original input embedding and the reconstructed one.

$$\mathcal{L}_{rec}\left(\mathbf{x}, \phi_1, \phi_2\right) = 1 - \cos\left(\mathbf{x}, d_{\phi_2}\left(\mathbf{z}\right)\right) \tag{4.3}$$

Overall, we aim to strike a balance between privacy protection (optimizing $\mathcal{L}_{disc}$, $\mathcal{L}_{adv}$) and utility (optimizing $\mathcal{L}_{rec}$) of the processed embeddings. The overall system is trained by alternating gradient descent steps on the parameters of the auto-encoder $\phi = \{\phi_1, \phi_2\}$ and the parameters of the discriminator θ:

$$\begin{aligned} \phi &\leftarrow \nabla_\phi \left(\mathcal{L}_{adv} + \mathcal{L}_{rec}\right) \\ \theta &\leftarrow \nabla_\theta \mathcal{L}_{disc} \end{aligned} \tag{4.4}$$

At test time, we produce a protected embedding $\tilde{\mathbf{x}}$ by passing $\mathbf{x}$ through the auto-encoder:

$$\tilde{\mathbf{x}} = d_{\phi_2}\left(e_{\phi_1}\left(\mathbf{x}\right)\right) \tag{4.5}$$

At this stage, the need for the discriminator module is no longer required, as our encoder-decoder network is assumed to be trained sufficiently to reconstruct speaker embeddings that effectively preserve identity while concealing gender-related information.

The evaluation of the capability of the Gender-AAE to preserve privacy involves assessing an attacker's ability to infer the gender of the original speaker from the protected utterance $\tilde{\mathbf{x}}$. To measure it, we train an external gender classifier $c(\cdot)$ on a separate set of clean embeddings, then report the gender classification performance of $c(\cdot)$ on the original test embeddings and their privacy-protected version: the difference between the two represents the effectiveness of gender concealment technique. The utility preservation is evaluated by comparing the performance of the same ASV system on the original and protected speaker embeddings.

We perform a preliminary evaluation of the reconstructed speaker embeddings of the Gender-AAE and obtain Area Under the ROC Curve (AUC) for gender classification = 98.45 (10^{-2}) and EER = 1.86% for ASV performance. In contrast to EER, AUC provides a comprehensive view, which is ideal for evaluating system security across diverse threshold selections. In order to ensure that the predictions of the gender classifier are truly random, the AUC must be close to 0.5. Therefore, it is necessary to strengthen the adversarial performance to conceal gender information.

In this work, we investigate the impact of adding noise derived from a Laplace mechanism, introduced in 2.1.2.2. Our choice of the Laplace mechanism is deliberate, as it is not only well-regarded for its noise addition and calibration properties but also because it offers robust DP guarantees. We have opted for the Laplace mechanism due to its suitability for addressing empirical queries, such as classification tasks, where the effectiveness of noise addition is crucial. Furthermore, the Laplace mechanism ensures that the latent vectors $\mathbf{z}$ maintain LDP guarantees, and the post-processing property of DP extends this guarantee to the reconstructed vectors, as discussed in Chapter 2.1.2.2.

4.2.2 Gender-Adversarial Auto-Encoder with Laplace noise

To enhance the ability of the Gender-AAE to conceal gender-related information and further strengthen the adversarial training of the encoder, we introduce a Laplace mechanism to the learned latent space. More specifically, during training, we incorporate a Laplace layer denoted as $dp(\cdot)$ into the process. This layer introduces a noisy vector $\mathbf{n} \sim \text{Laplace}(0, \Delta f/\epsilon)$ to the latent embedding $\mathbf{z}$. Figure 4.2 graphically depicts the system. In order to apply the Laplace mechanism effectively, we need to determine the sensitivity Δf. As outlined in Section 2.1.2.2, the concept of ℓ_1-sensitivity involves measuring how much the output of a function can change in response to a small change in its input. In our Gender-AAE framework, the function f refers to the encoding mechanism that transforms raw data into a latent representation. The ℓ_1-sensitivity, therefore, captures the maximum possible change in the encoded output when a single data point in the input is altered.

To calculate the ℓ_1-sensitivity for the Gender-AAE, we examine the encoder's

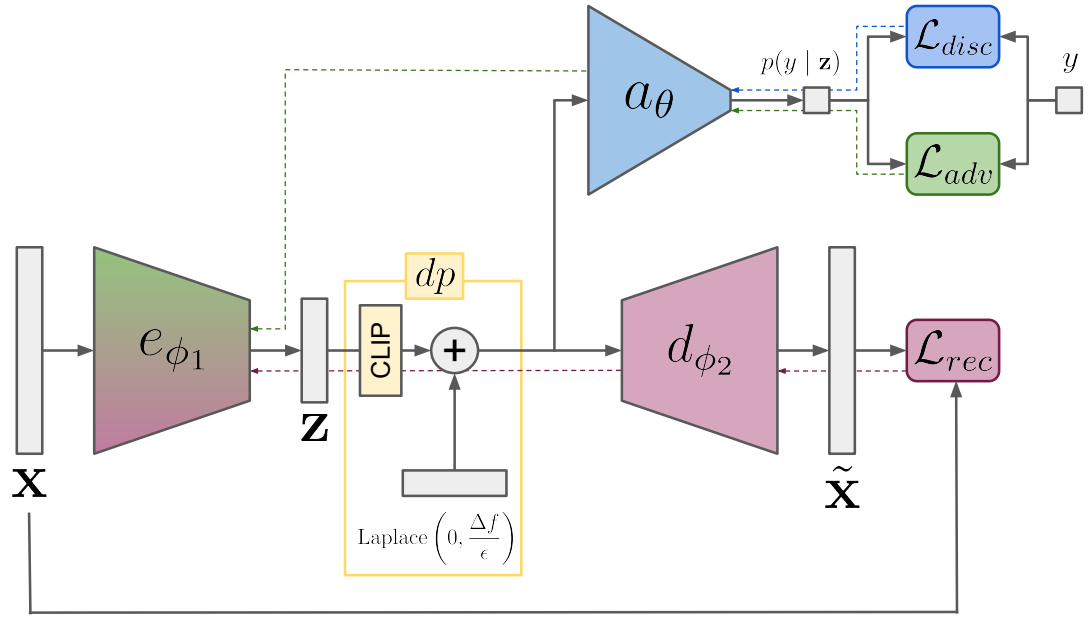


Figure 4.2: Illustration of the proposed system at training time. Solid and dashed arrows represent forward and backward propagation respectively. Modules are colored based on which gradient signal they are optimized by.

behavior with respect to two *adjacent inputs*, denoted by x and x'. In this context, adjacent inputs refer to pairs of data records or vectors that are part of the dataset $\mathcal{X}$ and are identical in every aspect except for one individual's data. This difference could be the presence or absence of an individual's record or a change in one individual's attributes while keeping the rest of the data unchanged. Formally, we calculate Δf it as:

$$\Delta f = \max_{x,x' \in \mathcal{X}} \| e_{\phi_1}(\mathbf{x}) - e_{\phi_1}(\mathbf{x}') \|_1 \tag{4.6}$$

However, since the e_{ϕ_1} function is the result of a neural network's encoding process, their values are not predetermined or bounded by any specific range. Each dimension of the latent vectors can take on a wide range of values, making it challenging to set an upper bound on their ℓ_1-norm. Therefore, we use the same clipping procedure used in [199]. By the triangle inequality for norms, we have:

$$\|e_{\phi_1}(x) - e_{\phi_1}(x')\|_1 = \|e_{\phi_1}(x) + (-e_{\phi_1}(x'))\|_1 \leq \|e_{\phi_1}(x)\|_1 + \|-e_{\phi_1}(x')\|_1 \tag{4.7}$$

Since the norm of a vector is equal to the norm of its negation, we get:

$$\|e_{\phi_1}(x)\|_1 + \|-e_{\phi_1}(x')\|_1 = \|e_{\phi_1}(x)\|_1 + \|e_{\phi_1}(x')\|_1 \tag{4.8}$$

Assuming that the ℓ_1-norm of $e_{\phi_1}(x)$ is constrained by the clipping threshold C for all $x \in \mathcal{X}$, it follows that:

$$\Delta f \leq \|e_{\phi_1}(x)\|_1 + \|e_{\phi_1}(x')\|_1 \leq C + C = 2C \tag{4.9}$$

Therefore, the ℓ_1-sensitivity Δf is bounded by $2C$. The latent space representation $\mathbf{z}$ is scaled by a coefficient $1/\max(1, \|\mathbf{z}\|_1/C)$. This method ensures that if $\|z\|_1 \leq C$, $\mathbf{z}$ remains unchanged, while if $\|z\|_1 > C$, it is scaled down to have a norm of C.

In practice, one pragmatic approach to determine an appropriate value for C is to compute the median of the norm of unclipped $\mathbf{z}$ vectors throughout the training phase. The value of the privacy budget ϵ can be chosen according to the desired balance between privacy protection and the utility of the produced embeddings.

The Laplace layer is then defined as

$$dp(\mathbf{z}) = \frac{\mathbf{z}}{\max\left(1, \frac{\|\mathbf{z}\|_1}{C}\right)} + \mathbf{n} \tag{4.10}$$

and has no learnable parameters. It is applied before $\mathbf{z}$ is passed to the decoder $d_{\phi_2}(\cdot)$ and to the discriminator $a_\theta(\cdot)$. The rest of the forward pass, the loss computation, and the overall training method then proceed as reported in Section 4.2.1. Once the model has been trained, the adversarial module $a_\theta(\cdot)$ is removed.

The integration of Laplace noise into the system fulfills a dual purpose. During the training phase, it acts as a regularizer for both the adversarial module and the decoder, thereby enhancing the ability of the Gender-AAE to obscure gender information. In the testing phase, the Laplace layer affords theoretical guarantees of privacy protection, as elucidated previously. A pivotal advantage of DP is its post-processing property, presented in Chapter 2.1.2.2. Similarly to the work in [200], we introduce noise into the latent space of the auto-encoder throughout the training process. Leveraging the post-processing attribute of differential privacy, we validate the privacy guarantees: the composition $d_{\phi_2} \circ dp$ satisfies ϵ-DP, thereby ensuring that the entire auto-encoder pipeline, denoted by $d_{\phi_2} \circ dp \circ e_{\phi_1}$, is also

compliant with ϵ-DP.

4.3 Experimental Evaluation and Results

In this section, we explore the experimental configurations, methodologies employed, and a detailed analysis of the results obtained. The core aim of our empirical study is to assess the effectiveness of the Gender-AAE in concealing gender information, with a specific focus on evaluating how the integration of the Laplace mechanism contributes to achieving our privacy objective. We investigate the utility of the speaker embeddings produced by the Gender-AAE in terms of speaker verification performance. We conduct a rigorous examination to determine whether the Laplace noise effectively enhances privacy without significantly compromising the utility of the reconstructed representations.

4.3.1 Databases

For our study, we used the VoxCeleb1 and VoxCeleb2 speaker recognition databases [201, 202]. VoxCeleb1 encompasses a substantial collection of over 100,000 utterances from 1,251 celebrities, while VoxCeleb2 is even more extensive, featuring over a million utterances from 6,112 speakers. Both datasets, compiled from YouTube videos, are extensively employed in the field of speaker recognition, as well as in various voice-related machine learning tasks.

System	Subset	Dataset	Number of utterances	
			Male	Female
Gender-AAE	Training	Subset of the VoxCeleb2 development set	397,032	397,032
External Gender Classifier	Training	Subset of the VoxCeleb1 development set	61,616	61,616
All systems	Evaluation	Subset of the VoxCeleb1 test set	2,900	2,900

Table 4.1: Statistics of the database used in training and evaluating the gender adversarial auto-encoder with and without the Laplace noise layer, as well as the external gender classifier

In our experimental setup, we used different datasets to train our systems, as presented in table 4.1. The Gender-AAE is trained on a carefully curated subset of the VoxCeleb2 development partition, containing 397,032 segments for each class

(male and female). Additionally, an external gender classifier is trained using a subset of the VoxCeleb1 development partition, comprising 61,616 segments for each gender class. For evaluation purposes, testing is conducted on a subset of the VoxCeleb1 test partition, consisting of 2,900 segments per class.

4.3.2 Experimental setting

In our experimental setup, the feature extraction for speaker embeddings is accomplished using ECAPA-TDNN [203], generating embeddings of dimension $d = 192$. The core components of our model include an encoder and decoder, both constructed as single-layer fully-connected NNs, and gender classifiers (comprising both the discriminator and an external classifier) built as two-layer fully-connected NNs.

The architecture of the encoder is augmented with a ReLU activation followed by batch normalization, while the decoder utilizes a tanh activation function. We design the latent space with a dimensionality of $l = 64$. The adversarial classifier within our model features a two-layer architecture: the first layer comprises 64 input units with ReLU activation, and the second layer consists of 32 units with a sigmoid activation function.

Furthermore, an external gender classifier is deployed, mirroring the architecture of the discriminator. It is intended for use by a hypothetical attacker aiming to infer gender, thereby allowing us to rigorously evaluate the efficacy of our privacy protection measures. This classifier is composed of 192 input units in its first layer and 100 units in the second layer.

For speaker verification, we adopt a methodology where a unique template is constructed for each speaker. Trial scores are subsequently generated by comparing trial embeddings with the corresponding speaker templates using cosine similarity. The training process is carried out with Adam optimizer using a learning rate of $1 \cdot 10^{-3}$ and a minibatch size of 128.

In the context of setting the clipping threshold C for our differential privacy mechanism, we determine its value by computing the median of the norms of all unclipped latent vectors z during training, resulting in $C = 18.35$.

To better explore the functional difference between the Laplace noise at training and at testing time, we perform experiments by independently varying the value of the privacy budget ϵ during the training phase (ϵ_{tr}) and during the test-

ing phase (ϵ_{ts}).
In the initial phase of our study, we train a series of models, each with a unique ϵ_{tr} value. This methodology enables a systematic examination of the privacy-utility trade-off, granting us an intricate understanding of how modifications in the privacy budget ϵ_{tr} influence the models' performance and robustness. Notably, a smaller ϵ_{tr} introduces more noise, thereby bolstering privacy but potentially disrupting model accuracy. Our aim is to delineate the extent to which the privacy budget ϵ_{tr} influences the overall performance and robustness of the models. Subsequently, in the testing phase, each model—characterized by its respective ϵ_{tr}—is rigorously evaluated. Here, a Laplace layer, parameterized by ϵ_{ts}, is integrated into the architecture, mirroring the approach during training. The models are individually assessed to discern the impact of the distinct ϵ_{tr} and ϵ_{ts} values on their efficacy, thus providing a comprehensive understanding of the dynamics between the training and testing phases in the context of privacy preservation and utility.

4.3.3 Gender-neutral speaker representation analysis

Our evaluation starts by analyzing models trained with various ϵ_{tr} values. Initially, we set the Laplace layer of the Gender-AAE with $\epsilon_{ts} = \infty$ during testing, which means removing the noise addition and retaining only the encoder-decoder architecture. This step is crucial to establish the best achievable model performance, serving as an upper bound for the system's utility, given that added noise typically reduces accuracy. As illustrated in Figure 4.3, the term ϵ_{tr} represents the specific epsilon value used during training. The figure presents the resulting ASV EER and gender classification AUC for each model, with each point representing a model trained with a distinct ϵ_{tr}. Notably, $\epsilon_{tr} = \infty$ implies the absence of DP protection during training, aligning closely with the performance of the original system before the integration of the Laplace layer. For every model, we analyze the output of the Gender-AAE, assessing the gender classification AUC using an external classifier and the ASV EER through cosine similarity. As anticipated, models with higher ϵ_{tr} values exhibit behavior akin to the original Gender-AAE system.

Our experimental approach fine-tunes the noise scale, focusing on regions where the privacy/utility trade-off is most pronounced and adopting a low resolution where variations are minimal. As observed, privacy and utility metrics tend to be

inversely correlated. Notably, an ϵ_{tr} value of 15 achieves a balanced compromise, yielding a gender classification AUC of 0.55 and an ASV EER of 8.1%. For comparison, the same gender classifier and ASV system score nearly 1 in AUC and 1.1% in EER, respectively, on the original ECAPA embeddings.

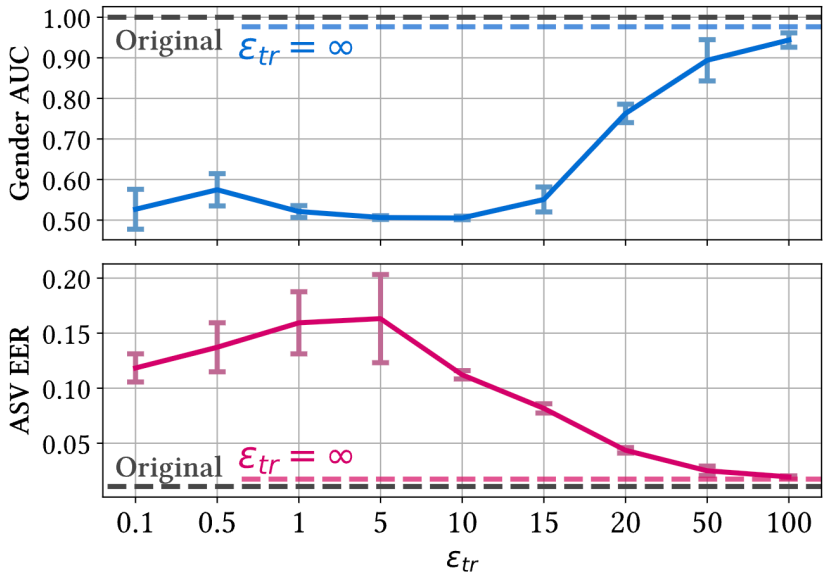


Figure 4.3: ASV EER and gender classification AUC achieved by the system for increasing values of ϵ_{tr}.

We select the model weights trained with $\epsilon_{tr} = 15$ and $\epsilon_{tr} = 20$, experimenting with $\epsilon_{ts} < \infty$ to integrate DP protection into the speaker embeddings, as depicted in Figure 4.4. Aligning ϵ_{ts} with ϵ_{tr} significantly bolsters gender concealment: AUC scores diminish to 0.50 from 0.55 and to 0.55 from 0.76 for $\epsilon_{tr} = \epsilon_{ts} = 15$ and $\epsilon_{tr} = \epsilon_{ts} = 20$ respectively. However, this adjustment results in an approximate 20-percentage point deterioration in ASV EER for both settings.

Incrementing ϵ_{ts} by 20 units effectively recuperates the ASV EER to about 10% for both model configurations, while maintaining favorable AUC scores of 0.55 and 0.68 for $\epsilon_{tr} = 15$ and $\epsilon_{tr} = 20$, respectively. These outcomes underscore the system's inherent flexibility post-training, concurrently ensuring DP protection of the generated embeddings.

Informal experiments with $\epsilon_{tr} = \infty$, corresponding to a training phase with no DP noise, resulted in embeddings that were not adequately protected. Even when applying a strict DP mechanism during the testing phase (using a low ϵ_{ts}

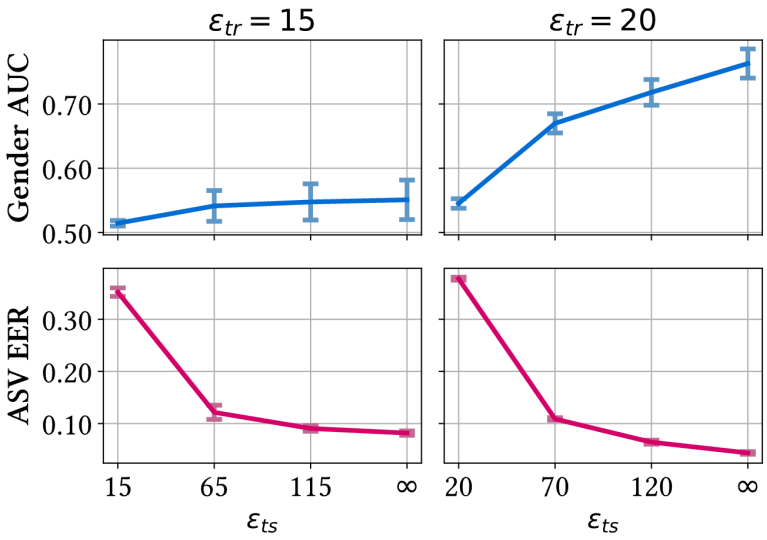


Figure 4.4: ASV EER and gender classification AUC achieved by the system for increasing values of ϵ_{ts}, for the cases of $\epsilon_{tr} = 15$ and $\epsilon_{tr} = 20$.

value), the privacy of the speaker embeddings was not sufficiently safeguarded and unuseful data for the speaker verification. This outcome emphasizes the necessity of integrating Laplace noise during the training phase to ensure that the privacy protection is effective during the testing phase as well.

4.4 Summary

This chapter has introduced an auto-encoder-based system that combines differential privacy mechanism with an adversarial auto-encoder for concealing gender-related information in speaker embeddings, while safeguarding their utility for speaker verification tasks. Traditionally used for data anonymization, the DP mechanism, in our innovative approach, collaborates with AAE to ensure individual identity in speaker verification is maintained, with a specific focus on masking gender details. The concealment of gender is performed through an adversarial game between an auto-encoder and an external gender classifier. Our improvement upon prior work includes the integration of a Laplace-noise-addition layer within the architecture. This inclusion not only regularizes the training phase to enable more robust concealment of gender but also equips the output speaker embeddings with solid DP guarantees at inference time. Fine-tuning the ϵ parameter

of the Laplace layer empowers our system to offer a customizable balance between privacy protection and utility, even after the model has been trained. The experimental results validate the effectiveness of our proposed solution in maintaining gender privacy while preserving the utility for speaker verification tasks. The ability of our approach to tailor the trade-off between privacy and utility makes it an adaptable and forward-thinking solution for privacy-preserving applications.

Chapter 5

Fairness and Privacy in Voice Biometrics: A Study of Gender Influences

In this chapter, we pursue the same motivation as the previous chapter (Chapter 4) for concealing gender attributes using an alternative approach. Additionally, we study the impact of the gender attributes on ASV systems performance. We evaluate as well the fairness of the systems (see Chapter 2.2.2) to ensure equitable outcomes for both male and female groups. This evaluation aligns with the GDPR principle of fairness discussed in Chapter 1.2. The organization of this chapter is as follows: It begins with a detailed exploration of the methodology for automatic speaker verification, gender recognition, and gender suppression using wav2vec 2.0. This section covers aspects such as pre-training and fine-tuning procedures. Subsequently, the experimental setup is described, including the databases used, metrics employed, fine-tuning procedures, and gender privacy threat models. The chapter then proceeds to present the experimental results, including assessments of utility, privacy, and fairness. Finally, the chapter concludes with a summary of of the main findings and contributions.

5.1 Automatic speaker verification, gender recognition and suppression using wav2vec 2.0

In this section, we outline our use of the wav2vec 2.0 model [35], a versatile speech feature encoder that is pre-trained through self-supervision and can be adapted to specific tasks. We fine-tuned wav2vec 2.0 for three distinct tasks: speaker verification, and gender recognition, and gender information suppression. Section 5.1.1 elaborates on the pre-training process of wav2vec 2.0, while Section 5.1.2 details our contributions to fine-tuning. Both procedures are graphically depicted in Fig. 5.1.

5.1.1 Pre-training

Given a raw audio input signal $\mathbf{x}$, wav2vec 2.0 processes it to generate a sequence of feature vectors $\mathbf{c}_1, \ldots, \mathbf{c}_T$. The model architecture consists of two principal components: a 1D-convolutional encoder and a Transformer module [204]. The encoder transforms the raw audio $\mathbf{x}$ into a series of latent feature vectors $\mathbf{z}_1, \ldots, \mathbf{z}_T$, which are then input into the Transformer module. Known for its effectiveness in modeling long-range dependencies in sequential data, the Transformer module outputs the final feature vectors $\mathbf{c}_1, \ldots, \mathbf{c}_T$ and concurrently generates quantized macro-codewords $\mathbf{q}_1, \ldots, \mathbf{q}_T$. These macro-codewords are composed by concatenating G codewords $\mathbf{q}_{t,1}, \ldots, \mathbf{q}_{t,G}$, each selected from distinct codebooks $\mathcal{Q}_1, \ldots, \mathcal{Q}_G$. These codebooks, repositories of vector representations, are thoroughly learned during training to capture specific data features or patterns. The selection of codewords from each codebook is directed by a probabilistic distribution, optimized during the model's pre-training phase. This distribution is computed as $\mathbf{p}_{t,j} = \mathrm{GS}(\mathbf{z}_t)$, where GS represents a linear projection of $\mathbf{z}_t$ to V dimensions, succeeded by a straight-through Gumbel-softmax estimator [205]. The Gumbel-softmax estimator allows the model to differentiably sample discrete codewords, a critical feature for the optimization of categorical distributions during training. This capability is particularly valuable when model decisions require interpretability or when the model needs to choose from a discrete set of options, such as selecting codewords from codebooks.

During the pre-training phase, wav2vec 2.0 aims to minimize two distinct loss functions: the *contrastive* loss $\mathcal{L}_m$ and the *diversity* loss $\mathcal{L}_d$. The contrastive loss is

instrumental in ensuring that the feature vector $\mathbf{c}_t$ generated by the Transformer aligns closely with its corresponding quantized macro-codeword $\mathbf{q}_t$, and simultaneously diverges from other non-corresponding macro-codewords within the batch. This loss is particularly vital when some of the latent feature vectors $\mathbf{z}_1, \ldots, \mathbf{z}_T$ are randomly masked, as it guides the Transformer to effectively recover the masked information by comparing $\mathbf{c}_t$ to the correct macro-codeword $\mathbf{q}_t$ and distinct distractor macro-codewords $\tilde{\mathbf{q}}$ sampled from the batch. The diversity loss $\mathcal{L}_d$, on the other hand, plays a crucial role in promoting the uniform use of all the V codewords within each codebook. It achieves this by maximizing the entropy of the average probability distribution $\bar{\mathbf{p}}_g$, computed from all $\mathbf{z}_t$ in a batch for each codebook g. This loss encourages the model to explore and use the full range of codewords, ensuring that the representations are diverse and comprehensive. The overall loss is articulated as follows:

$$\mathcal{L} = \underbrace{-\sum_{\substack{\text{masked} \\ \text{steps } t}} \log \frac{\exp\left(s(\mathbf{c}_t, \mathbf{q}_t)/\kappa\right)}{\sum_{\tilde{\mathbf{q}}} \exp\left(s(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa\right)}}_{\mathcal{L}_m} \underbrace{-\alpha \frac{1}{GV} \sum_{g=1}^{G} H\left(\bar{\mathbf{p}}_g\right)}_{\mathcal{L}_d} \tag{5.1}$$

In this equation, κ denotes the temperature coefficient, which modulates the sharpness of the softmax distribution, s represents the cosine similarity function, α is a hyperparameter that balances the influence of the two loss components, and H is the entropy, reflecting the diversity of codeword usage across the codebooks.

5.1.2 Fine-tuning for Speaker Verification and Gender Recognition

In this study, we fine-tune a wav2vec 2.0 model to adapt it for the specific downstream tasks of speaker verification and gender recognition. This fine-tuning process involves adjusting the pre-trained model parameters to enhance its performance on these specific tasks.

For each input utterance $\mathbf{x}$, the model generates a sequence of output features $\mathbf{c}_1, \ldots, \mathbf{c}_T$. These features are then aggregated across the temporal dimension to form a single 1-dimensional embedding $\mathbf{c}$. This embedding encapsulates the essential characteristics of the input utterance, serving as a distilled representation of the audio signal for subsequent processing.

In the context of gender recognition, the embedding $\mathbf{c}$ is forwarded through a

Figure 5.1: Graphical depiction of the proposed systems. M_s: fine-tuning the speaker identification task. Msg: fine-tuning gender and speaker identification. M_{sga}: similar to M_{sg}, but the gender identification task is made adversarial.

dedicated linear layer f_g. This layer is specifically trained to classify the gender by optimizing the cross-entropy loss $\mathcal{L}_g$. This loss function measures the discrepancy between the predicted logits, obtained from the linear layer, and the actual gender labels of the utterances, where '0' denotes male and '1' denotes female.

Conversely, for speaker verification, the embedding $\mathbf{c}$ is processed by a distinct linear layer f_s. This layer consists of N output neurons, with N representing the total number of speakers in the training dataset. The layer is intricately tuned to perform speaker identification by minimizing the additive angular margin (AAM) softmax loss $\mathcal{L}_s$ [206]. This loss function enhances the discriminative power of the model, ensuring that the embeddings of different speakers are well-separated in the embedding space. During the testing phase, the embedding $\mathbf{c}$ is used as a trial or enrollment vector, serving as the basis for verifying the identity of speakers.

The overall optimization objective of the model is defined by the following loss function:

$$\mathcal{L} = \lambda\mathcal{L}_s + (1 - \lambda)\mathcal{L}_g \tag{5.2}$$

where λ is a hyperparameter, ranging from 0 to 1, that balances the contribution of each task-specific loss component to the total loss.

In our experiments, we explore three distinct model configurations: Model 1 (M_s) is exclusively fine-tuned for speaker verification, implying $\lambda = 1$; Model 2 (M_{sg}) is fine-tuned for both tasks, setting $\lambda = 0.5$; Model 3 (M_{sga}) is similarly optimized for both tasks but incorporates a gradient reversal layer (GRL) g_r [207]. The GRL is a unique component that introduces an adversarial dynamic to the training process. Its purpose is to learn representations that are useful for the primary task (speaker verification) while being invariant to the secondary task (gender classification). During the forward pass, the GRL acts as an identity function, allowing the data to pass unchanged. However, during the backpropagation phase, it modifies the gradient by reversing its sign. As such, for any given loss that passes through the GRL, the gradient is multiplied by -1. Mathematically, if we consider the GRL as a pseudo-function $R(\mathbf{x}) = \mathbf{x}$ for the forward pass, its derivative with respect to $\mathbf{x}$ during the backward pass is $\frac{dR}{d\mathbf{x}} = -\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. This negation of the gradient effectively means that if the gender classification layer f_g is reducing the loss $\mathcal{L}_g$, the GRL forces the feature extractor

to maximize it instead, thereby obscuring gender-related features.

In this adversarial setup, while f_g seeks to minimize $\mathcal{L}_g$, the backbone of the model, preceding the GRL, learns to generate features that confuse the gender classifier. This is achieved by optimizing the model to maximize $\mathcal{L}_g$, effectively promoting the learning of gender-agnostic features. At the same time, the speaker verification task, which does not pass through the GRL, continues to minimize the speaker identification loss $\mathcal{L}_s$. The overall loss optimized during training can thus be formulated as usual (Equation 5.2).

5.2 Experimental setup

Described in this section are the databases used for all experimental work, the metrics used for evaluation, and the fine-tuning procedure.

Dataset	Usage	Number of speakers		Male-Female Imbalance (%)
		Male	Female	
VoxCeleb2 dev partition	Fine-turning	3682	2312	22.9
VoxCeleb1 test partition	Test	25	15	25

Table 5.1: Statistics of the datasets used for fine-tuning and evaluating the three models.

5.2.1 Databases

We used the VoxCeleb1 [201] and VoxCeleb2 [202] speaker recognition databases. Fine-tuning is performed using the VoxCeleb2 development set which contains data collected from 5994 unique speakers of which 3682 are male and 2312 are female, corresponding to an imbalance in favour of male speakers of 22.8% (61.4% and 38.6% female). To assess the performance of our systems, we used the VoxCeleb1 test set, which consists of 40 unique speakers of which 25 are male and 15 are female.

5.2.2 Metrics

To comprehensively evaluate the proposed models, a suite of metrics, primarily sourced from biometric classification systems such as speaker verification and

gender classification, was employed. These metrics facilitate a holistic assessment of utility, privacy, and fairness, three pivotal dimensions of system performance.

5.2.2.1 Utility and Privacy metrics

Utility is quantified by the performance of automatic speaker verification, specifically measured by the equal error rate.

Privacy is evaluated based on the challenge it presents to an adversary in inferring sensitive attributes from the model's output. To this end, the AUC is used as an indicator of the ability of the system to protect gender attributes against unauthorized detection.

5.2.2.2 Fairness Metrics

Fairness is assessed using two metrics: (i) fairness discrepancy rate and (ii) fairness activation discrepancy (FAD).

Fairness discrepancy rate examines the fairness of the outcome of the ASV system to ensure equitable treatment across different demographic groups. The FDR metric focuses on the balance between the false match rate (FMR) and the false non-match rate (FNMR) in assessing demographic-related differential performance [163,164]. It introduces two components: the false positive differential (FPD) and the false negative differential (FND). These differential terms represent the maximum discrepancy in FMR and FNMR, respectively, between any two demographic groups d_i and d_j, belonging to a set D, at a specific decision threshold τ. The FDR quantifies these discrepancies, modulated by risk parameters α and $1-\alpha$, to weight the relative importance of FMR and FNMR differences according to the security needs of a given application. High-risk situations, for instance, require a lower FMR to minimise security breaches. The FDR value ranges from 0 to 1, with 1 indicating full fairness. The calculation of the FDR proceeds as follows:

$$\mathrm{FPD}(\tau) = \max\left(\left|\mathrm{FMR}_{d_i}(\tau) - \mathrm{FMR}_{d_j}(\tau)\right|\right) \qquad \forall d_i, d_j \in D \tag{5.3}$$

$$\mathrm{FND}(\tau) = \max\left(\left|\mathrm{FNMR}_{d_i}(\tau) - \mathrm{FNMR}_{d_j}(\tau)\right|\right) \qquad \forall d_i, d_j \in D \tag{5.4}$$

$$\mathrm{FDR}(\tau, \alpha) = 1 - (\alpha \mathrm{FPD}(\tau) + (1-\alpha)\mathrm{FND}(\tau)) \tag{5.5}$$

In addition to the FDR, the area under FDR curve (auFDR) is used to compare the three ASV systems in terms of demographic differentials. The auFDR is determined by combining the FDR across a particular threshold range τ, denoted as FAR_x. To ensure a fair comparison of auFDR between various systems, it is necessary to report the range of τ used, as the value of the auFDR depends on this range. The auFDR spans from 0 to 1, with higher values indicating better fairness. In our experiments, we set the range for FARs from 0.001 to 0.1; FARs exceeding this threshold indicate a system of limited practical relevance.

Fairness Activation Discrepancy: is a metric derived from *InsideBias* [36], a fairness metric initially developed for face biometrics. The InsideBias metric identifies bias by analyzing layer activations and comparing of the responses of the model to demographic groups within different layers. Similarly, the FAD metric is used to study fairness within the network layers. The adaptation of FAD for voice biometrics is a novel metric in this context.

In voice biometrics, network outputs at each layer can be conceptualized as bi-dimensional tensors representing neurons across temporal frames:

$$A_{ij}^{[l]} = \Psi^{[l]}(\cdot) \tag{5.6}$$

Here, $i = 1, ..., N$ and $j = 1, ..., M$, where A_{ij} denotes the activation of the i^{th} neuron at the j^{th} temporal frame, $\Psi^{[l]}$ is the activation function at layer l, and N and M represent the total number of neurons and frames, respectively. For each layer l, we compute the root mean square of A_{ij} over the frames to capture significant positive or negative activations. We then determine the maximum value across the neuron dimension:

$$\Lambda^{[l]} = \max_i \sqrt{\left(\frac{1}{M}\sum_j A_{ij}^2\right)} \tag{5.7}$$

FAD is subsequently computed as the absolute difference in Λ between two distinct demographic groups, defined as $FAD = |\Lambda_{d_1} - \Lambda_{d_2}|$. Values of FAD approaching zero suggest greater fairness, indicating minimal discrepancy in activation intensities between the groups under comparison.

5.2.3 Fine-tuning procedure

As outlined in Section 5.1.2, the M_s, M_{sg}, and M_{sga} models undergo a fine-tuning process post-initialization. This process begins with a warm-up phase for the linear classification layers, which is conducted for the first $10,000$ optimization steps. During this phase, the wav2vec 2.0 backbone remains frozen, allowing the classification layers to adjust to the task-specific objectives without disturbing the pre-learned representations. The entire model, including the previously frozen backbone, is subjected to an end-to-end fine-tuning procedure. This comprehensive fine-tuning ensures that the entire model is optimally adjusted to the tasks at hand, allowing for a harmonious integration of the pre-learned representations with the new classification objectives.

The pre-trained wav2vec 2.0 model, provided by Baevski et al. [208][1], serves as the starting point for our fine-tuning process. The effectiveness of our fine-tuning strategy is evident from the performance metrics: for the speaker identification task, all three models attained an accuracy exceeding 95%, showcasing their robustness in identifying speakers. In contrast, the adversarial model, M_{sga}, achieved a gender recognition accuracy of only 47%. This significantly reduced accuracy for gender recognition underscores the efficacy of the adversarial approach in obfuscating gender-related features, thereby enhancing the privacy aspects of the model without compromising its ability to perform speaker verification.

5.2.4 Gender privacy threat models

To assess to ability of the ASV systems to conceal gender information in speaker embeddings, we simulate a third party, namely an *attacker*, by training a 2-layer fully-connected NN model ($\mathscr{N}$). This model predicts the gender of the speaker from the embeddings generated by the the fine-tuned wav2vec 2.0. We examine two attack scenarios:

1. *Uninformed attack* (uIA): the attacker is unaware of any gender concealment techniques. Thus, the attacker trains the gender classifier $\mathscr{N}$ using embeddings that gender-protected, which are generated by the M_s and M_{sg} models.

[1] `https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec`

2. *Informed attack* (IA): the attacker knows that a specific model (M_{sga}) has been used to protect gender identity and has access to it. The attacker then trains $\mathscr{N}$ using embeddings produced by the protected model M_{sga}. We expect this scenario to result in a more effective attack as the attacker understands the protection mechanism.

5.3 Experimental results

Results are reported for the three models M_s, M_{sg}, and M_{sga}. Performance is assessed in terms of utility, privacy, and fairness.

5.3.1 Utility

In terms of utility, results show that speaker verification capabilities of the M_s model, which is fine-tuned solely for speaker verification, aligns with state-of-the-art ASV systems [1, 7], achieving an EER of 2.36%, as presented in Table 5.2. On the other hand, models M_{sg} and M_{sga}, which incorporate a focus on gender attributes, demonstrate a slightly inferior performance achieving an EER of 3.23% and 3.89% respectively. This suggests that gender-related factors do not contribute significantly to enhancing speaker verification. Moreover, a further examination of the EER based on gender, reveals minimal disparities in speaker verification between the two genders.

		Models		
		M_s	M_{sg}	M_{sga}
EER(%)	Overall	2.36	3.23	3.89
	Male	3.12	4.22	4.98
	Female	3.05	4.21	5.26

Table 5.2: Performance analysis of the three models for utility, including EER breakdown by gender

5.3.2 Privacy

Privacy performances are presented in Table 5.3.

AUC results for uninformed attacks are presented at the top. When the gender classifier $\mathscr{N}$ is trained and tested using unprotected speaker embeddings extracted with the M_s and M_{sg} models,the AUC is 97.09% and 98.07% for M_s and M_{sg},

	Data		Attack
	Training	Test	AUC (%)
uIA	M_s	M_s	97.09
	M_s	M_{sga}	**46.80**
	M_{sg}	M_{sg}	98.07
	M_{sg}	M_{sga}	**40.76**
IA	M_{sga}	M_{sga}	96.27

Table 5.3: Assessment of gender concealment effectiveness under different threat scenarios in terms of AUC.

respectively. This indicates a lack of privacy protection. In contrast, when training the gender classifier with unprotected embeddings and test with gender-protected embeddings provided by the M_{sga} model, the AUC drops to 46.80% and 40.76% for M_s and M_{sg}, respectively. The notable decrease in AUC suggests that the gender classifier predictions approach randomness, effectively obscuring gender information and thereby demonstrating a successful privacy protection measure.

The performance results of the informed attack are shown in the last row of Table 5.3. When embeddings are derived using the M_{sga} model, the AUC notably rises to 96.27%. This underscores the challenge of concealing gender information from embeddings.

Fig. 5.2 offers insight into this matter. It presents a visualization generated by PCA (explained in Section 2.1.2.3) of the embeddings produced by each of

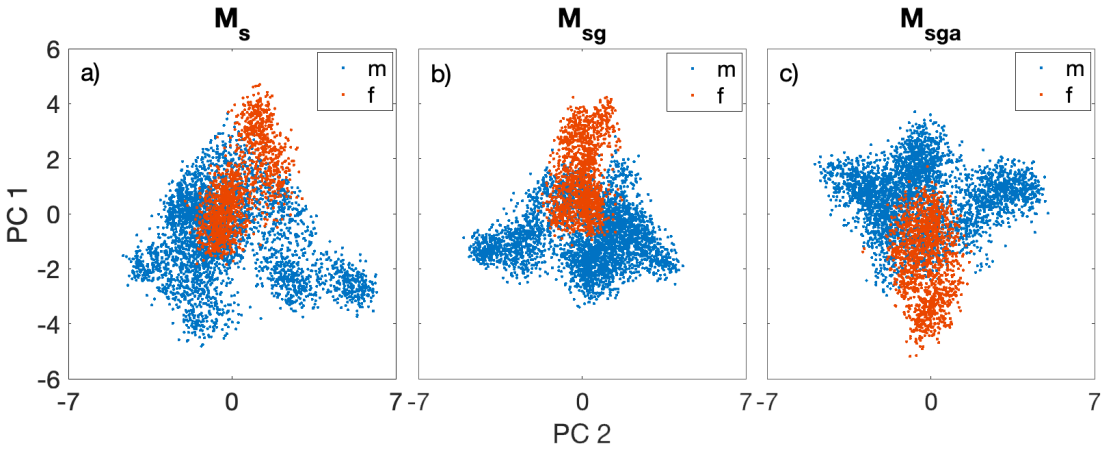


Figure 5.2: PCA visualizations of features from three models illustrating gender recognition capabilities. Blue points correspond to males and red to females.

the three models. Despite the M_{sga} model being trained to mitigate gender cues, Fig. 5.2c demonstrates their persistence. Instead of completely obscuring gender cues, M_{sga} merely rotates the principal components, which explains why gender recognition remains feasible even when trained on similarly processed training data. Despite the adversarial training of the M_{sga} model to suppress gender cues, as illustrated in Fig. 5.2c, these factors persist. It's apparent that rather than fully disentangling gender cues, the M_{sga} model only rotates the principal components. This is why, when trained on similarly pre-processed training data, gender can still be identified.

5.3.3 Fairness

Fairness assessment is performed using FDR, auFDR, and FAD. Results of the auFDR for different values of α are shown in Table 5.4. The auFDR results of the three models are close to 1, indicating reasonable fairness for each group.

In Figure 5.3, we present a graph illustrating the FDR for all three systems across various thresholds, with $\alpha = 0.5$. Notably, the FDR consistently exceeds 0.9 for all cases, with the M_s system consistently ranking as the fairest for each threshold (τ). Once more, it is evident that gender influence fails to enhance fairness.

			Models		
			M_s	M_{sg}	M_{sga}
auFDR	α	0	0.98	0.97	0.96
		0.25	0.97	0.97	0.95
		0.5	0.97	0.96	0.94
		0.75	0.96	0.95	0.92
		1	0.95	0.94	0.91

Table 5.4: Performance analysis of auFDR across various α values (refer to eq.5.5) for τ ranging from 0.1% to 10%.

. Results of the assessment of the internal bias of the three models is depicted in Figure 5.4. The FAD metric has been performed at different network layers of each model while considering two groups: male and female. This analysis seeks to offer insights into how fairness is measured across the three models and how these FAD measurements change across different layers of the network. By examining the internal bias at each layer, we aim to gain a clearer understanding of how

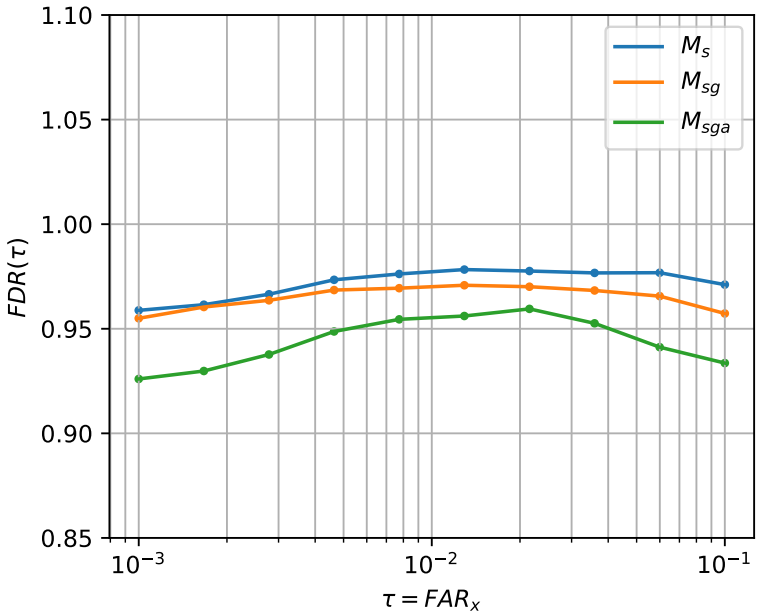


Figure 5.3: FDR of different ASV systems for different decision thresholds for τ from 0.1% to 10%

factors such as model architecture and training data influence fairness outcomes.

As depicted in Fig. 5.4, a total of 32 layers were selected from the wav2vec 2.0 model. Among these, 8 layers originate from the 1D-convolutional encoder, while the remaining 24 layers stem from the Transformer modules.

Fig. 5.4 illustrates the FAD values computed across various layers. The first layers of the convolutional neural networks (CNNs) display comparable fairness, possibly due to their focus on low-level features. On the other hand, Transformer layers, responsible for processing high-level features, exhibit wider variations in fairness. The M_s and M_{sga} models demonstrate complementary behavior: when one model achieves high FAD, the other tends to have lower FAD, and vice versa. This observation could stem from the fact that the M_s model was fine-tuned for speaker verification, whereas M_{sga}, equipped with a gradient reversal layer, aimed to suppress gender information.

As layers progress, all models eventually converge to FAD values, with M_s emerging as the fairest model by the end, aligning with observations regarding auFDR fairness measures.

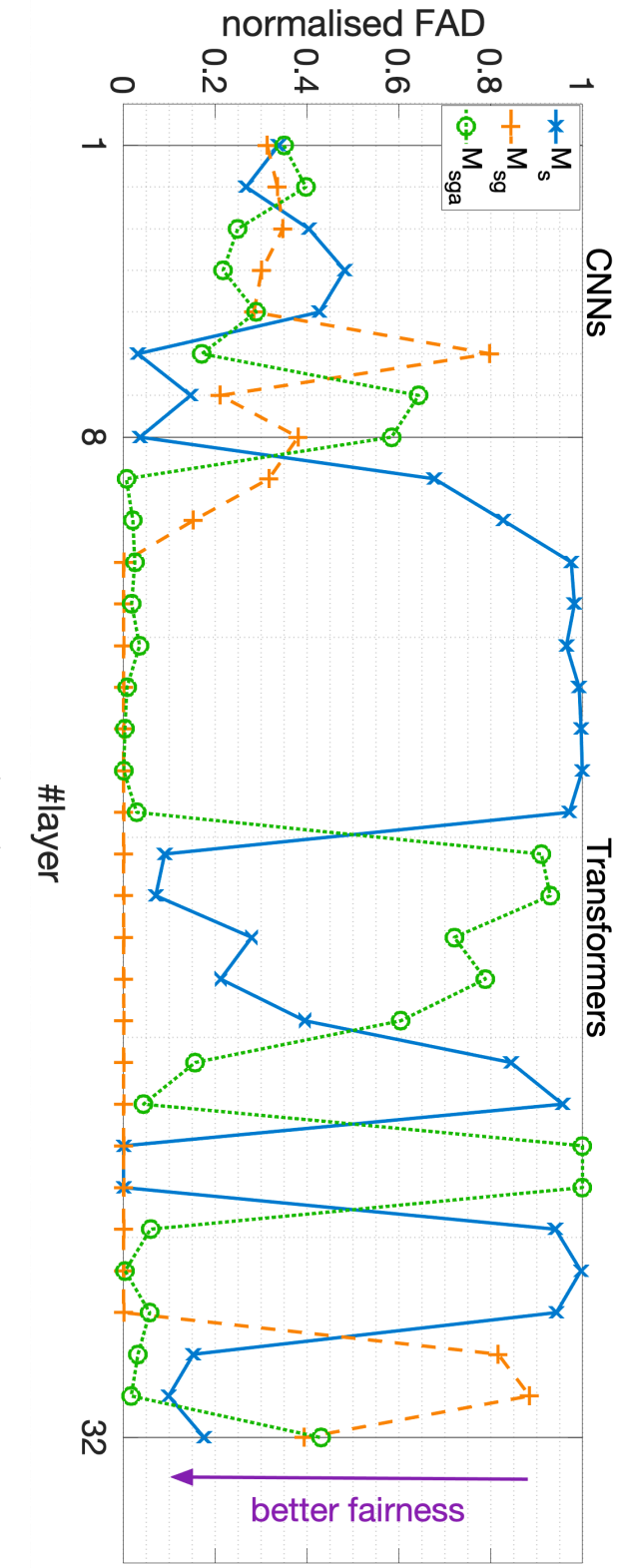


Figure 5.4: Normalised Fairness Activation Discrepancy (FAD) of different systems at different wav2vec 2.0 module layers.

5.4 Summary

In this chapter, we have introduced a study focusing on the influence of gender information during the fine-tuning process of wav2vec 2.0 for speaker verification. We have proposed three models: M_s, M_{sg}, and M_{sga}, each with distinct objectives: speaker recognition, speaker recognition with gender identification, and speaker recognition with gender concealment, respectively.

Our experiments reveal that M_s achieves successful speaker verification (with an EER of 2.36%), while M_{sga}, designed to conceal gender information, performed less effectively (with an EER of 3.89%). Interestingly, enhancing gender recognition within the M_{sg} model does not improve speaker verification performance (with an EER of 3.23%). Privacy assessments indicate effective gender concealment against uninformed attacks, although informed attackers could still extract gender information. Fairness evaluations, based on FDR, show that highlighting or concealing gender do not notably affect the fairness of the systems. Additionally, we have introduced the fairness activation discrepancy metric tailored for speech data as a method for analyzing network fairness. An analysis of FAD across model layers demonstrates more discrepancies within Transformer layers, but eventually, all systems converge to FAD values consistent with the auFDR assessment, with system M_s displaying superior fairness.

To summarize, while we achieve notable results in utility and privacy protection against uninformed attacks, future work should focus on strengthening gender concealment against informed attacks and enhancing fairness across systems.

Chapter 6

A Comparison of Differential Performance Metrics for the Evaluation of Automatic Speaker Verification Fairness

Fairness is a crucial aspect in the development and deployment of biometric systems, ensuring equitable treatment across various demographic groups. However, automatic speaker verification systems, despite their effectiveness, encounter fairness issues, as highlighted in Chapter 1.3. Adding to this challenge is the absence of an international standard for measuring fairness in biometric systems. Additionally, the majority of existing research in this domain has predominantly focused on face recognition, with limited attention given to voice recognition.

In this chapter, we aim to bridge this gap by directing our focus towards examining fairness metrics within the context of ASV systems. First, we present three candidate fairness metrics and fairness criteria in biometric recognition systems. We then detail the experimental setup, outlining the ASV systems used for generating outcomes to assess fairness metrics, the databases utilized, and the fairness evaluation procedure. Next, we present the findings of our experiments and engage in discussions regarding the behavior of fairness metrics and their alignment with the required criteria. Additionally, we evaluate the fairness of ASV systems using the most appropriate fairness metric. Finally, we conclude this chapter with a summary of our findings and insights.

6.1 Fairness Metrics and Criteria

In this section, we present three fairness metrics initially proposed for evaluating face recognition systems, coupled with essential criteria that such metrics must meet to effectively assess fairness. These guidelines ensure that the metrics possess both theoretical robustness and practical relevance for real-world applications.

6.1.1 Fairness Discrepancy Rate

The fairness discrepancy rate metric, previously introduced in Section 5.2.2.2, evaluates demographic disparities by considering false match and false non-match rates [163,164]. It quantifies the maximum differences in FMR and FNMR between demographic groups d_i and d_j at threshold τ, with FPD and FND components. FDR values range from 0 to 1, indicating fairness. Below is a reminder of the FDR equations:

$$\mathrm{FPD}(\tau) = \max\left(\left|\mathrm{FMR}_{d_i}(\tau) - \mathrm{FMR}_{d_j}(\tau)\right|\right) \qquad \forall d_i, d_j \in D \tag{6.1}$$

$$\mathrm{FND}(\tau) = \max\left(\left|\mathrm{FNMR}_{d_i}(\tau) - \mathrm{FNMR}_{d_j}(\tau)\right|\right) \qquad \forall d_i, d_j \in D \tag{6.2}$$

$$\mathrm{FDR}(\tau, \alpha) = 1 - (\alpha \mathrm{FPD}(\tau) + (1-\alpha)\mathrm{FND}(\tau)) \tag{6.3}$$

6.1.2 Inequity Rate

The inequity rate (IR) assesses fairness by examining the ratio of the highest and lowest FMR and FNMR values among different demographic groups d_i and d_j [164]. This is achieved by comparing the maximum FMR and FNMR with the minimum FMR and FNMR across all groups. Risk parameters α and $1-\alpha$ are once again employed to scale the ratios before their aggregation. Importantly, unlike the FDR, lower IR values indicate greater fairness. The IR is computed as follows:

$$\mathrm{FPD}(\tau) = \frac{\max_{d_i} \mathrm{FMR}_{d_i}(\tau)}{\min_{d_j} \mathrm{FMR}_{d_j}(\tau)} \qquad \forall d_i, d_j \in D \tag{6.4}$$

$$\mathrm{FND}(\tau) = \frac{\max_{d_i} \mathrm{FNMR}_{d_i}(\tau)}{\min_{d_j} \mathrm{FNMR}_{d_j}(\tau)} \qquad \forall d_i, d_j \in D \tag{6.5}$$

$$\mathrm{IR}(\tau, \alpha) = \mathrm{FPD}(\tau)^{\alpha} \cdot \mathrm{FND}(\tau)^{(1-\alpha)} \tag{6.6}$$

6.1.3 The Gini Aggregation Rate for Biometric Equitability

The Gini aggregation rate for biometric equitability (GARBE) is based on the Gini index, a measure of inequality [164, 209]. This metric uses a normalized Gini coefficient for n demographic groups. Normalization by $\frac{n}{n-1}$, as proposed in [210], corrects for the downward bias when the number of samples (demographic groups) is small.

The Gini coefficient associated with the FMR is defined as follows:

$$G_{\mathrm{FMR}}(\tau) = \frac{n}{n-1}\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|\mathrm{FMR}_{d_i}(\tau) - \mathrm{FMR}_{d_j}(\tau)|}{2n^2\overline{\mathrm{FMR}(\tau)}}\right) \tag{6.7}$$

where $\overline{\mathrm{FMR}}$ is the mean value.

Similarly, the Gini coefficient related to the FNMR is defined by:

$$G_{\mathrm{FNMR}}(\tau) = \frac{n}{n-1}\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|\mathrm{FNMR}_{d_i}(\tau) - \mathrm{FNMR}_{d_j}(\tau)|}{2n^2\overline{\mathrm{FNMR}(\tau)}}\right) \tag{6.8}$$

for any $d_i, d_j \in D$.

In adapting to be consistent with the notation above, the pair of Gini coefficients are combined according to:

$$\mathrm{FPD}(\tau) = G_{\mathrm{FMR}}, \quad \mathrm{FND}(\tau) = G_{\mathrm{FNMR}} \tag{6.9}$$

$$\mathrm{GARBE}(\tau, \alpha) = \alpha\mathrm{FPD}(\tau) + (1-\alpha)\mathrm{FND}(\tau) \tag{6.10}$$

GARBE values range between 0 and 1, where 0 signifies complete fairness and 1 denotes complete unfairness.

6.1.4 Functional Fairness Measure Criteria

The primary objective of fairness metrics is to assess and determine the most equitable classification algorithms. Howard et al. [209] outline the essential attributes required for such metrics, referred to as the functional fairness measure criteria (FFMC). These criteria emphasize the interpretability and practicality of the metrics:

1. **FFMC.1**: The contributions of FMR and FNMR to the fairness metric should be intuitive across typical risk parameters and operationally relevant

error rates.

2. **FFMC.2**: The metric must have well-defined boundaries, with minimum and maximum values, to establish clear benchmarks.

3. **FFMC.3**: The metric must remain computable even for demographic groups with no observed errors, which is increasingly common with the advancement of more accurate biometric algorithms.

6.2 Experimental Setup

In this section, we outline the ASV systems used for evaluating the fairness metrics, the database employed, and the procedure for fairness evaluation.

6.2.1 Speaker Verification Systems

We employ five distinct ASV systems for assessing fairness metrics, each featuring unique structural and functional characteristics.

1. The ECAPA system [203] utilizes a standard ECAPA-TDNN [1] architecture, integrating 3 SE-Res2Block modules to derive a 192-dimensional speaker embedding. It employs cosine similarity as its backend.

2. ResNetSE34L [211] is a streamlined version of ResNet-34 [8], employing self-attentive pooling (SAP) [212] to aggregate frame-level features into utterance-level features, focusing on the most informative frames. It uses squared Euclidean distance as a distance metric.

3. ResNetSE34V2 [213] is a performance-optimized variant of ResNet-34. The stride is removed at the first convolutional layer to reduce computational cost. It adopts attentive statistics pooling (ASP) [214] for temporal frames aggregation.

4. ERes2Net [2] improves upon the Res2Net structure by integrating local and global feature fusion, capturing both detailed and holistic patterns in the input signal.

5. CAM++ [3] mainly consists of a front-end convolution module and a densely connected time delay neural network (D-TDNN) backbone. It incorporates

an improved context-aware masking (CAM) module in each D-TDNN layer and employs multi-granularity pooling to capture discriminative speaker characteristics.

6.2.2 Databases

The pre-trained models[1,2,3] of the ASV systems used in our experiments are trained using the development set of the VoxCeleb2 database [202]. Table 6.1 shows at the top the statistics of the training subset. Although VoxCeleb2 includes multiple languages, the predominance of English speakers, and the English language as well, results in significant imbalance.

Subset	Dataset	# of speakers	# of nationalities	# of utterances
Training	VoxCeleb2 Dev	5,994	less or equal to 145*	1,092,009
Evaluation	VoxCeleb1	72	9	1728 (24 per speaker)

Table 6.1: Statistics of Datasets for training the five automatic speaker verification systems and evaluating the three fairness metrics.
*Only the total number of nationalities across the entire VoxCeleb2 dataset (combining both dev and test partitions) has been reported. The specific number of nationalities within each partition has not been provided

All evaluations are conducted using the combined VoxCeleb1 development and test sets [201]. The statistics for the evaluation subset are presented at the bottom of Table 6.1. To assess the utility of the five ASV systems, we established a balanced protocol. This protocol involved selecting speakers from nine different nationalities: *USA, UK, Germany, Australia, Italy, India, Ireland, New Zealand, and Canada*. From each nationality group, eight speakers were randomly chosen, resulting in a total of 72 speakers. For each speaker, 24 utterances were selected. The ASV protocol [4] of the pooled is composed of a total of 39,744 comparison trials. These trials are evenly distributed, comprising 2,208 mated and 2,208 non-mated combinations for each nationality.

[1] `https://github.com/TaoRuijie/ECAPA-TDNN`
[2] `https://github.com/clovaai/voxceleb_trainer`
[3] `https://github.com/alibaba-damo-academy/3D-Speaker/tree/3dspeaker`
[4] `https://github.com/OubaidaOubaida/FairnessMetricsEvaluation/blob/main/pooled_data.txt`

6.2.3 Fairness evaluation procedure

In this section, we introduce the evaluation procedure of the fairness metrics and the utility assessment of the ASV systems. Table 6.2 presents results from a preliminary analysis of the ASV performance in terms of pooled EER and FMR/FNMR at the threshold corresponding to the pooled EER, across nine groups based on different nationalities.

As expected, pooled EER results reveal differing performance levels across various ASV systems. The comparison also highlights consistent variations in the FMR and FNMR across nationality groups. For instance, results for the ERes2Net the outcomes for the ERes2Net system are most favorable for the UK group, exhibiting low FMR and FNMR. In contrast, although the same system demonstrates comparable security for the German group (with similar FMR), it lacks the same level of convenience (resulting in a higher FNMR) comparing to other groups. Interestingly, the opposite behaviour is observed for the Indian group. Similar diverging results are observed for other nationality groups. This analysis underscores the importance of a *single* measure which reflects fairness across the *full set* of groups. This approach is essential to guaranteeing the fairness of ASV systems, ensuring they do not unfairly disadvantage any specific group due to nationality or other demographic factors.

Noting that the EER is not suited to the assessment of *any* binary classifier in the case that a particular application calls for the prioritisation of a lower rate of FMR or FNMR [163, 175], it is similarly unsuitable as a measure of fairness. Another measure proposed by Toussaint et al. [173], previously presented in Section 2.2.3, is based on the min DCF. This metric also does not generalize and only considers one particular operating point.

Given that using different operating points inherently involves a trade-off between the FMR and FNMR, and as advocated in [163], any fairness metric must consider disparities in both and take them into account. While averaging FMR and FNMR rates across groups is possible, it results in two metrics that still require additional interpretation to serve as fairness indicators. This precisely aligns with what each of the three candidate metrics delivers.

Our approach to evaluating the proposed fairness metrics aligns with the methodology outlined in [209]. Initially, we adopt a benchmark threshold corresponding to an FMR of 0.1% for the initial assessment. Moreover, expanding

Nationality	ASV Systems											
	ERes2Net		***CAM++***		***ECAPA***		***ResNetSE34V2***		***ResNetSE34L***	
	FMR (%)	FNMR (%)	FMR (%)	FNMR (%)	FMR (%)	FNMR (%)	FMR (%)	FNMR (%)	FMR (%)	FNMR (%)
USA	1.22	1.04	1.40	1.54	1.13	1.45	1.36	1.68	2.76	2.08
UK	0.68	0.45	0.41	0.14	0.72	0.23	0.50	0.50	2.08	1.90
Germany	0.59	2.81	0.63	4.26	2.49	6.34	2.31	6.34	4.17	8.11
Australia	0.68	0.27	1.99	0.14	1.90	0.27	0.86	0.77	1.54	1.72
Italy	1.95	2.58	1.40	2.72	2.58	3.85	3.26	2.54	3.53	4.08
India	2.31	0.09	1.90	0.14	2.76	1.00	6.11	0.00	5.53	0.09
Ireland	0.18	2.04	0.82	2.31	0.77	1.45	0.45	2.36	0.86	4.21
New_Zealand	1.86	0.27	1.86	0.18	2.17	0.27	0.95	1.18	1.86	2.67
Canada	1.13	1.31	1.40	0.86	1.31	1.22	1.13	1.63	5.30	2.17
Pooled EER (%)	**1.18**		**1.37**		**1.79**		**1.88**		**3.01**	

Table 6.2: ASV performance in terms of pooled EER and FMR/FNMR at the threshold corresponding to the pooled EER, across nine groups of different nationalities. Color background transitions from better (green) to mid (yellow) to lower (red) performances.

upon the framework in [209], we adhere to the guidelines set forth in ISO/IEC DIS 19795-10 [164] by exploring a range of thresholds, specifically within an FMR range of 0.1% to 10%, along with a comprehensive risk parameter range spanning from 0 to 1. The rationale behind selecting a 0.1% to 10% FMR range is twofold: first, today's most advanced ASV systems achieve acceptable levels of FNMR at FMRs in the order of 0.1%; second, ASV systems (or any other biometric system) with an FMR exceeding 10% may have limited practical utility. Our objective is to examine variations in fairness across a range of representative operating points for five distinct systems, thus acquiring a comprehensive understanding of metric behavior. This analysis is crucial because, in real-world scenarios, each system operates optimally at a threshold tailored to the specific application it serves.

6.3 Experimental results and discussion

In this section we present an assessment of the fairness metrics presented in Section 6.1. First, we present the assessment results for a fixed threshold which produces an FMR = 0.1%. Second, we show the assessment results for a range of thresholds from 0.1% to 10%. Last, we evaluate each fairness metric in terms of the three FFMCs described in Section 6.1.4.

6.3.1 Metrics evaluation results at a fixed threshold

We assess the performance of fairness metrics across five distinct ASV systems. We aggregate the evaluation results to gain insights into how these metric behave across different systems.

For each metric, we present three plots. The first plot depicts the density distribution of the metric when the importance of the FMN and FNMR differentials are equal ($\alpha = 0.5$), similar to the approach outlined in the reference work [209]. The second plot showcases the density distribution of the metric for all alpha values within the range [0,1]. These two plots serve to illustrate the spread of metric values, aiding in determining whether the metric offers intuitive comparisons between systems and how it responds to varying risk parameter alpha. The third graph displays the FPD and FND terms for alpha values within the range [0,1] to study the scales of the error rates.

To ensure consistency with the methodology in [209], we evaluate the fairness metric at a fixed threshold. Experiments were conducted with decision thresholds

configured to produce an FMR of 0.1%.

6.3.1.1 FDR evaluation

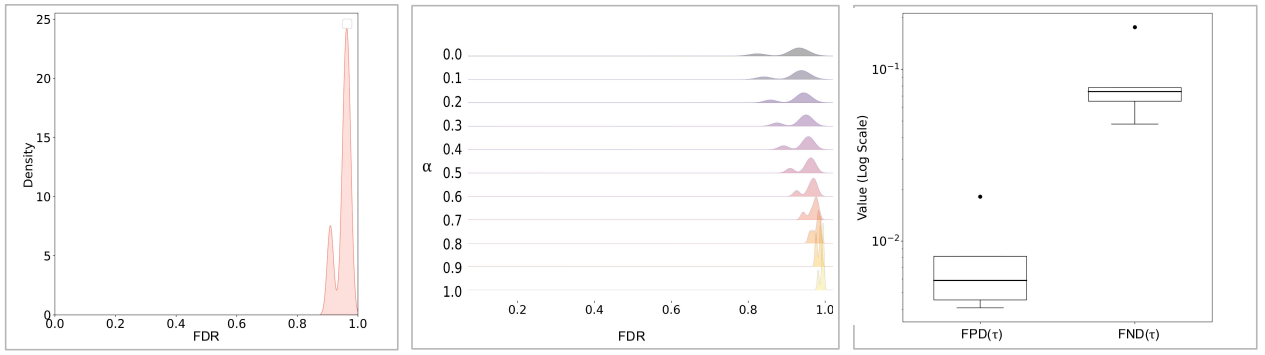
We start our evaluation by computing the FDR metric, described in Section 6.1.1, for the five ASV. The FDR values are depicted in Figure 6.1. In Figure 6.1(a) and 6.1(b), FDR values predominantly fall within the range [0.82-1]. Comparing the fairness of systems and gauging the influence of the risk parameter α, particularly in cases where systems are mostly fair, poses a challenge due to this concentration. It is challenging to intuitively determine which system is fairer and to evaluate the influence of the risk parameter α, particularly in the case of mostly fair systems. Figure 6.1(c) demonstrates that the differential terms (FPD, associated with the FMR, and FND, associated with the FNMR) operate on markedly different scales. Aggregating these terms poses a challenge in accurately configuring them with the term α. This complexity highlights the difficulty in intuitively assessing the contributions of FMR and FNMR within the FDR metric. As a result, this method does not meet the criteria set forth by the first FFCM principle (Section 6.1.4).

6.3.1.2 IR evaluation

The assessment conducted on the IR metric presented in Section 6.1.2 revealed instances where certain subgroups provided minFMR values of 0. This renders the computation of the FPD term in Equation (6.4) unfeasible, thereby resulting in the inability to compute the IR. Among the five evaluated ASV systems, the IR metric is only computable for the ResNetSEV2 system, with the value being 13.35. This observation underscores the incapacity of the IR metric to satisfy the third FFCM (Section 6.1.4). Consequently, this raises concerns about the suitability of IR as a reliable metric for fairness assessment in such contexts. Moreover, the ratio-based nature of the IR introduces an additional layer of complexity. Its values possess no upper limit, implying significant potential for variation and the possibility of extremely high values. This further complicates its interpretation.

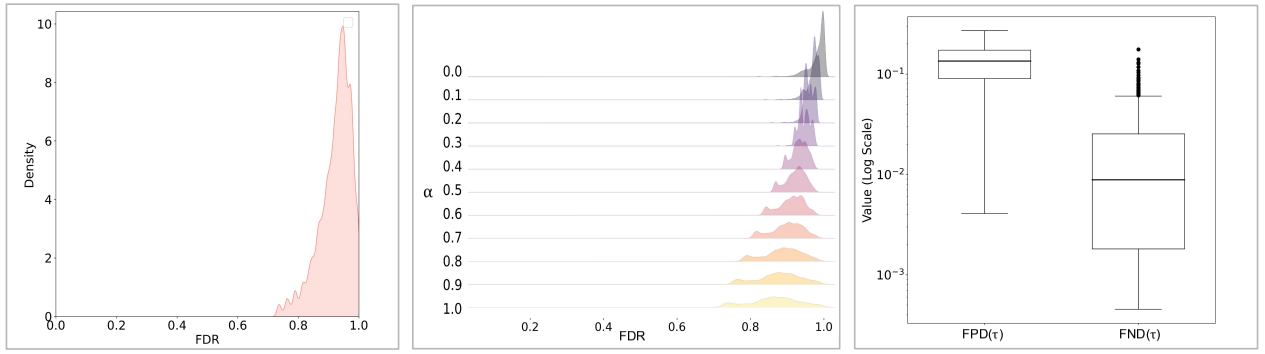
6.3.1.3 GARBE evaluation

We now shift our focus to the GARBE metric detailed in Section 6.1.3. The results presented in Figures 6.3(a) and 6.3(b) show a broader range compared to the FDR values. This range spans from 0.19 to 0.61 for α in the interval



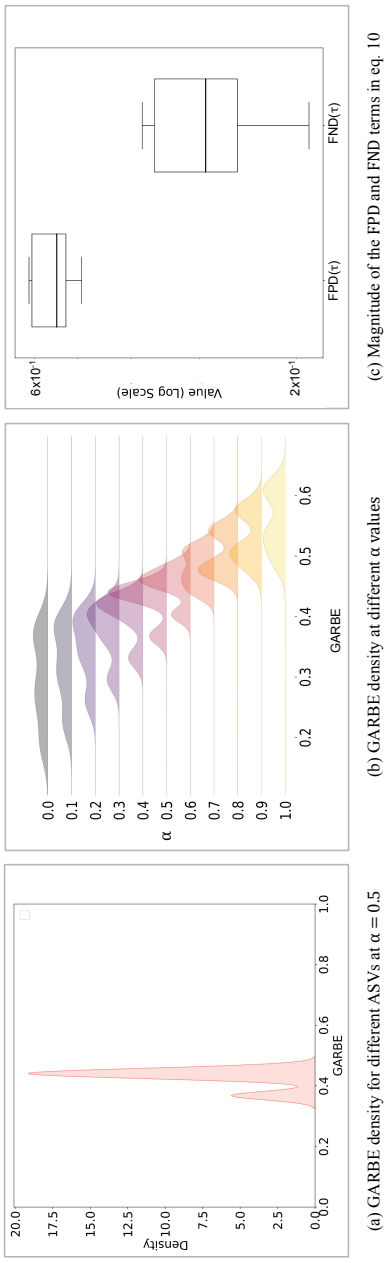
(a) FDR density for different ASVs at $\alpha = 0.5$ (b) FDR density at different α values (c) Magnitude of the FPD and FND terms in eq. 3

Figure 6.1: FDR values using 5 automatic speaker verification systems at a threshold corresponding to FMR = 0.1%



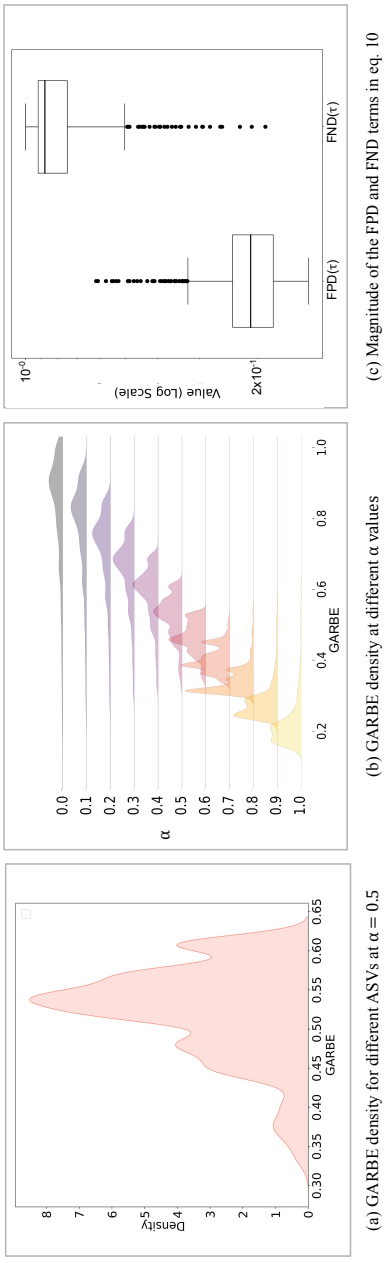
(a) FDR density for different ASVs at $\alpha = 0.5$ (b) FDR density at different α values (c) Magnitude of the FPD and FND terms in eq. 3

Figure 6.2: FDR values using 5 automatic speaker verification systems at a range of thresholds corresponding to a FMR varying from 0.1% to 10%



(a) GARBE density for different ASVs at $\alpha = 0.5$ (b) GARBE density at different α values (c) Magnitude of the FPD and FND terms in eq. 10

Figure 6.3: GARBE values using 5 automatic speaker verification systems at a threshold corresponding to FMR = 0.1%



(a) GARBE density for different ASVs at $\alpha = 0.5$ (b) GARBE density at different α values (c) Magnitude of the FPD and FND terms in eq. 10

Figure 6.4: GARBE values using 5 automatic speaker verification systems at a range of thresholds corresponding to a FMR varying from 0.1% to 10%

[0,1], which is approximately half of the theoretical range. This provides a more intuitive comparison between systems and a better assessment of the impact of α. An additional critical finding is related to the FPD and FND terms, shown in Figure 6.3(c). These terms are scaled to a comparable magnitude. Specifically, the median value for the FPD term is found to be 0.55, while the median for the FND term is observed at 0.29. The normalization of the Gini coefficient computation reduced the discrepancy in the scale of the differential terms. As a result, the impact of α becomes more pronounced.

6.3.2 Metrics evaluation results at different thresholds

In an extension of the previous study, we broaden the assessment scope to encompass not only various systems but also different operational points. We extend the range of our analysis by adjusting the threshold of the five ASV systems, spanning from a 0.1% to 10% FMR. This methodology ensures that our evaluation captures scenarios reflective of real-world conditions.

For each metric, we provide three plots as in the previous section. The only distinction this time is that the FMR values in the plots range from 0.1% to 10%.

6.3.2.1 FDR evaluation

Despite adjusting both the threshold and α values, the FDR values exhibit a consistent trend, with values concentrated between 0.72 and 1, as depicted in Figures 6.2(a) and 6.2(b). Comparing which system is fairer remains non-intuitive, as most FDR values nearly overlap for all α values. Additionally, the persistent scale disparity between the FPD and FND terms, illustrated in Figure 6.2(c), suggests challenges in intuitively understanding the contributions of FMR and FNMR to the FDR metric. Therefore, the FDR metric fails to meet the criteria outlined in FFMC.1.

6.3.2.2 IR evaluation

The results for the IR metric reaffirm the challenges associated with its computation in specific scenarios, even with a variable threshold range. The criterion FFMC.3 remains unfulfilled. The selective representation of 8.6% of computable values in Figures 6.5(a) and 6.5(b) illustrates the extensive range of the IR, which can reach up to 200 in certain instances. This underscores the unbounded nature

of the IR metric, rendering it non-compliant with FFMC.2. However, the FPD and FND terms exhibit a similar scale, as depicted in Figure 6.5(c). The use of a ratio-based approach ensures a more balanced comparison between terms, thereby fulfilling FFMC.1.

6.3.2.3 GARBE evaluation

The GARBE metric effectively overcomes the limitations observed in previous metrics. As depicted in Figures 6.4(a) and 6.4(b), GARBE values consistently span the entire theoretical range from 0 to 1. Moreover, Figure 6.4(b) illustrates the sensitivity of GARBE to changes in α. The FPD and FND terms, showcased in Figure 6.3(c), with median values of 0.21 and 0.88 respectively, are on the same scale. This ensures that both terms contribute significantly to the fairness assessment, thereby meeting FFMC.1.

Analysis of the boxplots reveals a swap in the positions of the FPD and FND terms between Figures 6.1(c) and 6.2(c), as well as between 6.3(c) and 6.4(c). This swap occurs because, at certain thresholds, the FPD either exceeds or falls below the FND, and vice versa. Specifically, in the analysis of the FDR metric, for thresholds resulting in an FMR lower than 0.9%, the FND term surpasses the FPD term. Conversely, for thresholds leading to an FMR above 0.9%, the FPD term becomes higher. Regarding the GARBE metric, thresholds that produce an FMR below 0.4% result in a higher FPD than FND. This explains the observed variation in the positions of the boxplots when the FMR is set at 0.1% and for FMR ranges from 0.1% to 10%.

6.3.3 Summary of the Fairness Metrics Criteria

Our assessment of fairness metrics for ASV within the framework of the Functional Fairness Measure Criteria (Section 6.1.4) reveals diverse findings regarding the FDR, IR, and GARBE metrics. These results are summarised in Table 6.3.

FFMC Criteria	FDR	IR	GARBE
FFMC.1		✓	✓
FFMC.2	✓		✓
FFMC.3	✓		✓

Table 6.3: Summary of Fairness Measures Criteria for ASV

(a) IR density for different ASVs at $\alpha = 0.5$

(b) IR density at different α values

(c) Magnitude of the FPD and FND terms in eq. 6

Figure 6.5: IR values using 5 automatic speaker verification systems at a range of thresholds corresponding to a FMR varying form 0.1% to 10%

The FDR metric inherently meets the FFMC.2 criterion as it is bounded, offering an interpretable measure of fairness where a value of 1 represents perfect fairness and a value of 0 signifies complete unfairness. Moreover, the FDR metric remains computable even when FNMR or FMR is zero, aligning with FFMC.3. However, challenges arise with FFMC.1 due to the disparate scales of the FPD and FND terms when using typical risk parameter ranges (α in [0,1]) and operationally relevant error rates (FMR in [0.1%,10%]). This discrepancy complicates the interpretation of the contributions of FMR and FNMR in the computation of the FDR metric. Thus, the FDR metric does not meet FFMC.1.

The IR metric satisfies FFMC.1 by adopting a ratio-based approach, which effectively balances the contributions of the FPD and FND terms. However, it encounters limitations in meeting FFMC.2 and FFMC.3 due to its unbounded nature. This characteristic makes it challenging to establish benchmarks and renders it incalculable when FNMR or FMR reach zero. Therefore, while the IR metric addresses one criterion, it falls short in fulfilling the others.

GARBE emerges as the most robust metric, satisfying all FFMC criteria. By leveraging the Gini coefficient, as outlined in Equations 6.7 and 6.8, the FPD and FND terms are converted to a same scale before their aggregation. This normalization is key for meeting FFMC.1 It ensures an intuitive understanding of the contributions of FMR and FNMR to the GARBE metric calculation, facilitating a nuanced and balanced representation across varying α values.

Moreover, GARBE fulfills FFMC.2 by maintaining set boundaries, enabling the establishment of clear benchmarks. Additionally, it remains computable even when error rates are zero, thereby meeting FFMC.3. Thus, GARBE not only addresses the fairness criteria comprehensively but also ensures practicality and interpretability in real-world scenarios.

Our analysis validates and extends the findings of the study on fairness for face recognition reported in [209]. It reinforces the conclusion that the GARBE metric is particularly well-suited for evaluating fairness in biometric systems. This consistency demonstrates that GARBE is not limited to face recognition systems but extends to studies of fairness in automatic speaker verification as well.

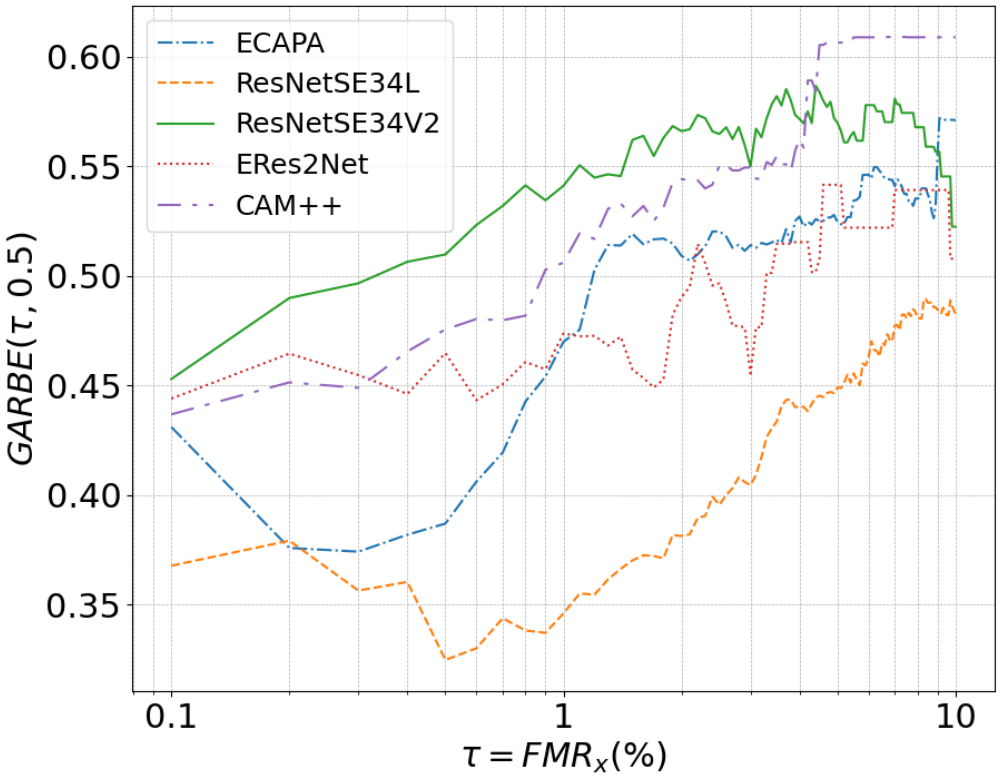


Figure 6.6: GARBE of different ASV systems for different decision $\tau = \text{FMR}_x$ where x varies from 0.1% to 10% and for $\alpha = 0.5$.

6.4 Fairness and ASV assessment

Although not the primary focus of this study, we now turn our attention to an analysis of the ASV systems themselves. We present an evaluation of the five ASV systems based on their verification performance (FMR vs. FNMR) and fairness, here assessed solely using the GARBE metric.

Figure 6.6 displays a plot of GARBE values over an FMR range from 0.1% to 10%, with $\alpha = 0.5$. Figure 6.7 presents a detection error trade-off (DET) plot. The ResNetSE34L system exhibits the lowest GARBE values across all FMR thresholds, suggesting that it is the fairest system. However, in terms of verification performance, the same ResNetSE34L system performs the poorest. In this case, enhanced fairness, characterized by lower differential performance across

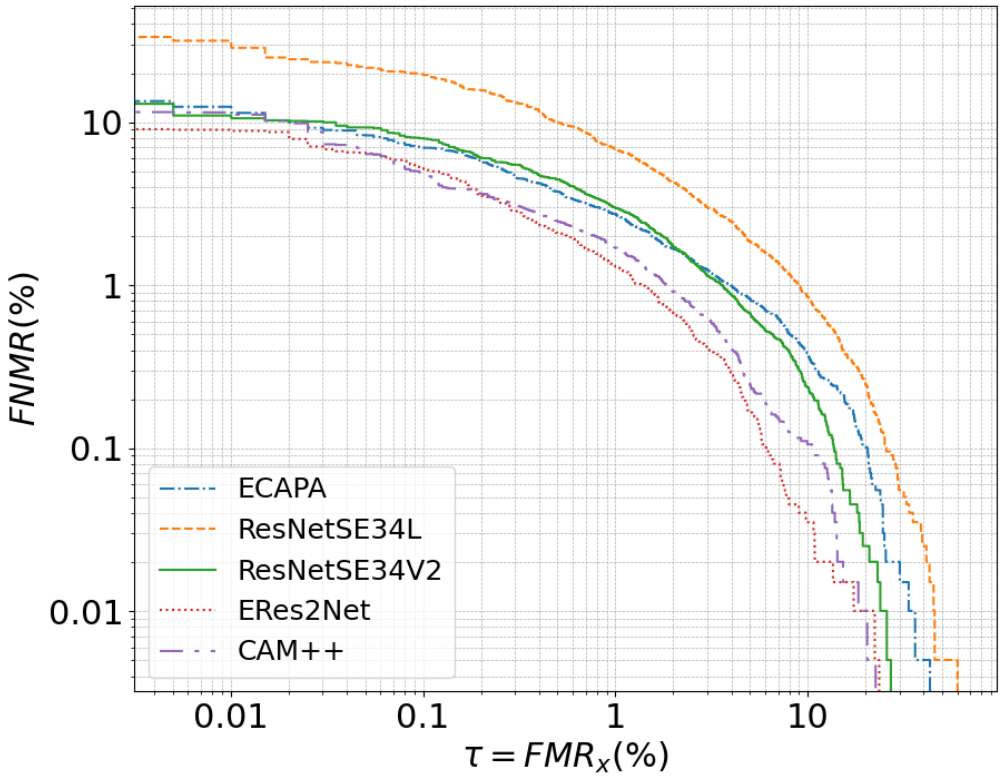


Figure 6.7: Detection error tradeoff (DET) curve of different ASV systems.

groups, comes at the expense of degraded verification performance.

GARBE values for the least fair system, ResNetSE34V2, are consistently highest across most operating points, indicating significant differential performance across groups. However, despite its fairness shortcomings, verification accuracy surpasses that of the ResNetSE34L system, suggesting an apparent trade-off between performance and fairness.

On the other hand, the ERes2Net and CAM++ systems, while not the fairest, particularly for FMRs under 1%, demonstrate comparable fairness levels. Additionally, they are among the top performers in terms of verification accuracy.

The ECAPA model displays distinct behavior compared to other systems. While its verification performance is average, there is considerable variation in GARBE values between lower and higher FMRs. This pronounced fluctuation

suggests a particular sensitivity unique to the ECAPA model, highlighting a distinct aspect of its operational characteristics.

6.5 Conclusions

In this chapter, we have conducted a comparison of three distinct fairness metrics, fairness discrepancy rate (FDR), inequity rate (IR), and Gini aggregation rate for biometric equitability (GARBE), within the context of automatic speaker verification. We have further performed an analysis of fairness and verification performance for five state-of-the-art ASV systems. Our findings indicate that the GARBE metric emerges as the most adept in meeting the Functional Fairness Measure Criteria (FFMCs).

Our analysis reveals a delicate balance between fairness and accuracy. The system deemed fairest exhibits the poorest verification performance, while the system with the highest verification accuracy demonstrates only average fairness. These findings underscore the challenge of achieving a balance between fairness and verification performance.

Given the requirement for fairness, the evaluation of fairness should be incorporated into the development process of ASV systems, just as it should be for any biometric system. Relying solely on raw verification performance does not ensure the creation of equitable solutions. This highlights the potential need for implementing *fairness by design*.

Chapter 7

Conclusions and Future Research

In this chapter, we provide a summary of the research conducted in this thesis. We begin by outlining the contributions and findings in Section 7.1, followed by a discussion of potential directions for future research in Section 7.2.

7.1 Summary

In this thesis, we considered the problem of enhancing compliance with European General Data Protection Regulation (GDPR) principles concerning data privacy and fairness in voice biometrics applications. We explored privacy concerns within the context of both automatic speaker verification (ASV) and countermeasure (CMs) systems. We further evaluated fairness of ASV systems and promoted a concept of *fairness by design*. Furthermore, in order to enhance the reproducibility of our research and allow comparisons with alternative approaches, we have employed common evaluation protocols and publicly available databases in our experimental assessments. We provide below a summary of the contributions and chapters of the thesis.

Chapter 2 provides a background and literature review on privacy and fairness enhancing technologies used for biometric systems.

Chapter 3 presents, PRIVASP, the first privacy preservation scheme for CMs systems using secure multi-party computation. PRIVASP not only preserves the privacy of CMs systems but also protects their intellectual property (IP) by keeping the model parameters private. Most state-of-the-art cryptographic-based privacy-preserving schemes often introduce system utility degradation or/and computational overhead compared to non-protected systems. The proposed privacy-

preserving CMs system successfully meet privacy requirements while maintaining reasonable spoofing detection performance. Following the *privacy by design* principle of the GDPR, a shallow neural network is designed from scratch to meet secure multi-party computation requirements. Two scenarios were considered depending on whether or not the CMs system provider wishes to keep the parameter of the model private to protect the IP. Experiments conducted on the ASVspoof 2019 Logical Access (LA) database validate the effectiveness of PRIVASP in real-time spoofing detection, as shown in Table 7.1, while operating with maintaining utility, as indicated in Table 7.2

system / type	**PRIVASP-1024**	**PRIVASP-512**	B01	B02	LFCC-GMM	RawNet2	ResNet18-SP
plaintext	2.8	2.7	339.9	89.9	100.6	12.0	2.8
scenario 1	95.8	59.9	-	-	-	-	-
scenario 2	349.6	208.1	-	-	-	-	-

Table 7.1: Average inference time in ms per utterance.

system	type	EER [%]	min-tDCF
B01	plaintext	9.57	0.2366
B02	plaintext	8.09	0.2116
LFCC-GMM	plaintext	3.50	0.0904
RawNet2	plaintext	5.54	0.1547
ResNet18-SP	plaintext	6.82	0.1140
PRIVASP-1024	plaintext	7.03	0.1485
	scenario 1	7.02	0.1481
	scenario 2	7.02	0.1481
PRIVASP-512	plaintext	7.10	0.1549
	scenario 1	7.13	0.1550
	scenario 2	7.13	0.1550

Table 7.2: Performance for the ASVspoof 2019 LA evaluation partition in terms of pooled EER and min t-DCF for the two baselines, B01 and B02, the high-spectral-resolution LFCC, RawNet2, ResNet18-SP and our proposed PRIVASP-1024 and PRIVASP-512 systems. PRIVASP systems are also evaluated in privacy-preserving scenario 1 and 2.

The results highlights the remarkable efficiency of PRIVASP in the traditionally resource-intensive realm of secure multi-party computation. Through the strategic design of a shallow neural network, PRIVASP seamlessly integrates with

the secure 2PC framework. This approach effectively reduces the computational overhead typically associated with such methods and maintains utility in both non-private and private domains.

Chapter 4 presented an innovative auto-encoder-based system that merges a differential privacy (DP) mechanism with an adversarial auto-encoder (AAE) to conceal gender-related information within speaker embeddings, while maintaining their utility for speaker verification purposes. The concealment process is performed through an adversarial game between the auto-encoder and an external gender classifier. A Laplace-noise-addition layer is integrated within the architecture to enhance the robustness in gender concealment during training and solidifying DP guarantees at inference time. The ability to fine-tune the Laplace noise by adjusting the privacy budget ϵ enables our system to provide a customizable balance between privacy protection and utility, even post-training.

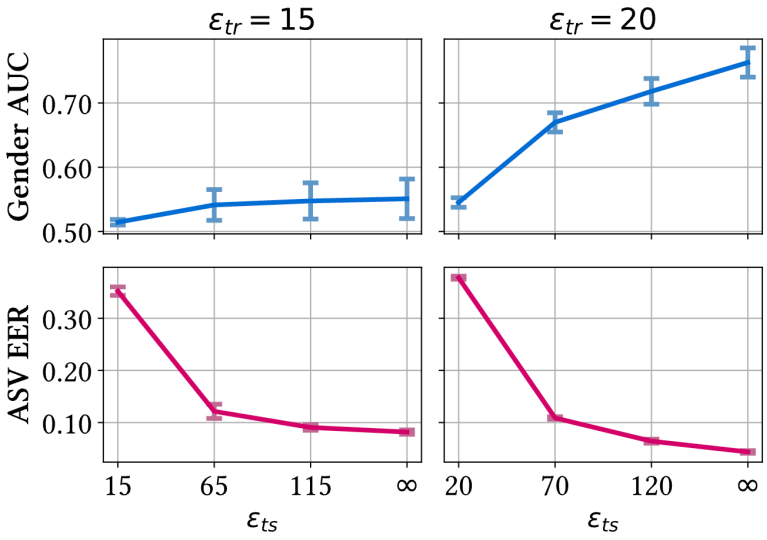


Figure 7.1: ASV EER and gender classification AUC achieved by the system for increasing values of ϵ_{ts}, for the cases of $\epsilon_{tr} = 15$ and $\epsilon_{tr} = 20$.

Experimental evaluations conducted on the VoxCeleb dataset demonstrate the effectiveness of our approach in carrying out speaker verification tasks while concealing speaker gender and maintaining ϵ-differential privacy guarantees (Figure 7.1).

In **Chapter 5**, we conducted a study on the impact of gender information during the fine-tuning of wav2vec 2.0 for speaker verification. Three models were

			Models		
			M_s	M_{sg}	M_{sga}
EER(%)	Overall		2.36	3.23	3.89
	Male		3.12	4.22	4.98
	Female		3.05	4.21	5.26
auFDR	α	0	0.98	0.97	0.96
		0.25	0.97	0.97	0.95
		0.5	0.97	0.96	0.94
		0.75	0.96	0.95	0.92
		1	0.95	0.94	0.91

Table 7.3: Performance analysis of the three models for utility and fairness, including EER breakdown by gender and auFDR across various α values (refer to eq.6.3) for τ ranging from 0.1% to 10%.

	Data		Attack
	Training	Test	AUC (%)
uIA	M_s	M_s	97.09
	M_s	M_{sga}	**46.80**
	M_{sg}	M_{sg}	98.07
	M_{sg}	M_{sga}	**40.76**
IA	M_{sga}	M_{sga}	96.27

Table 7.4: Assessment of gender concealment effectiveness under different threat scenarios in terms of AUC.

introduced: M_s, M_{sg}, and M_{sga}, each with distinct objectives: speaker recognition, speaker recognition with gender identification, and speaker recognition with gender concealment, respectively.

Experiments performed on the VoxCeleb dataset reveal that while M_s achieved successful speaker verification, M_{sga}, designed for gender concealment, performed less effectively. Surprisingly, enhancing gender recognition within M_{sg} did not improve speaker verification performance (Table 7.3). Privacy assessments indicated effective gender concealment against uninformed attacks, though informed attackers could still extract gender information (Table 7.4). Fairness evaluations based on auFDR (in Table 7.3) show that highlighting or concealing gender did not notably affect system fairness.

We also introduced the fairness activation discrepancy (FAD) metric tailored for speech data, revealing more discrepancies within Transformer layers, as de-

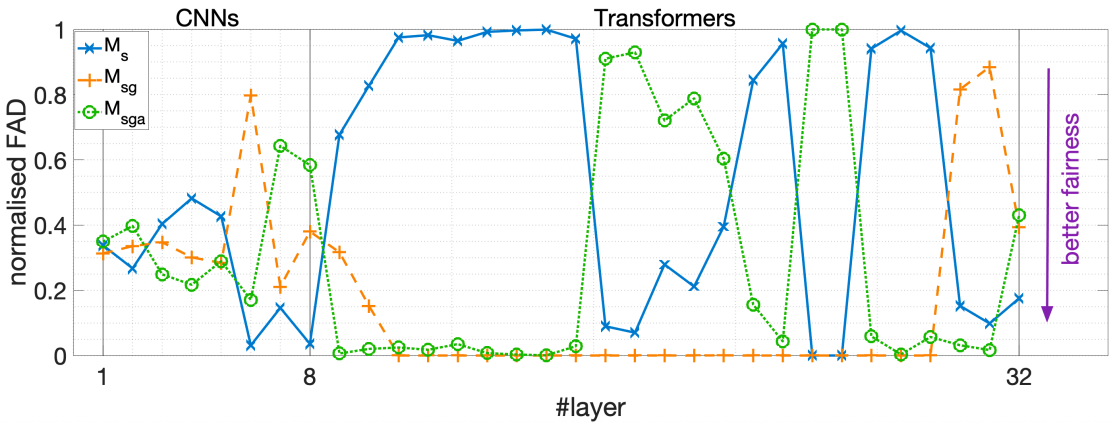


Figure 7.2: Normalised Fairness Activation Discrepancy (FAD) of different systems at different wav2vec 2.0 module layers.

picted in Figure 7.2. However, all systems eventually converged to consistent fairness values, with M_s displaying superior fairness.

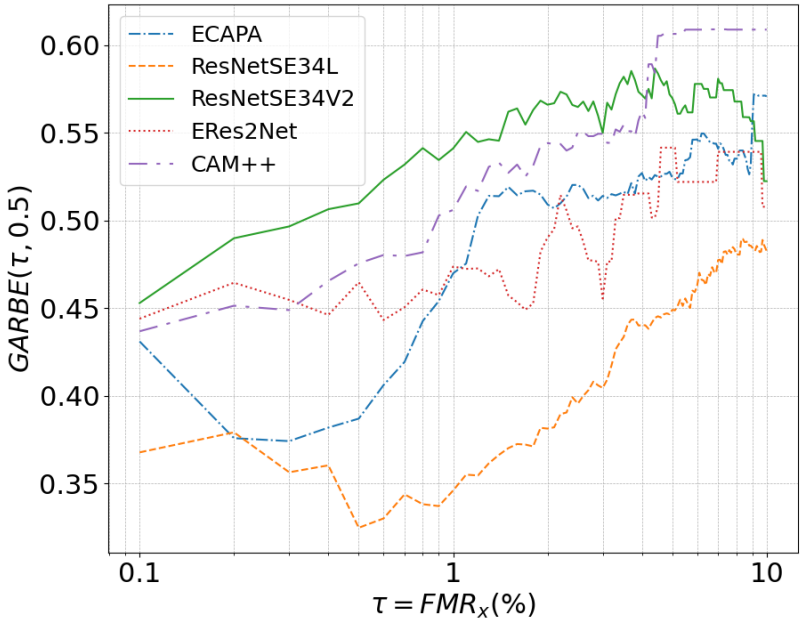


Figure 7.3: GARBE of different ASV systems for different decision $\tau = \text{FMR}_x$ where x varies from 0.1% to 10% and for $\alpha = 0.5$.

In **Chapter 6**, we addressed the challenge of the non-existent international

standards for fairness evaluation in the field of biometrics. We assessed three fairness measures-fairness discrepancy rate (FDR), inequity rate (IR), and Gini aggregation rate for biometric equitability (GARBE)-in the context of speaker verification and evaluate ASV systems using the most suitable fairness metric.

Our comparative analysis of three fairness metrics using the VoxCeleb dataset five state-of-the-art ASV systems reveals that the GARBE metric stands out as the most adept in meeting the Functional Fairness Measure Criteria (FFMCs).

We further used the GARBE metric to evaluate fairness of the five ASV systems. Experimental results provide insights into the delicate balance between fairness and accuracy. Despite the system with the highest verification accuracy demonstrating only average fairness, the one deemed fairest exhibits the poorest verification performance (Figure 7.3).

These findings stress the need to integrate fairness evaluation into ASV system development, emphasizing the need for implementing *fairness by design*.

7.2 Future Research Directions

Based on the findings presented in this thesis, the following potential directions for future research are identified:

- **Exploration of Disentanglement Techniques:** Future research directions in the realm of disentanglement techniques for ASV systems hold promise for advancing privacy preservation while retaining identity. Expanding on the disentanglement technique introduced in Chapter 4, which employs adversarial auto-encoder (AAE) architecture with a differential privacy (DP) mechanism to conceal gender information. Further enhancements could involve disentangling multiple soft biometric attributes simultaneously, such as age, accent, and emotion.

  Additionally, diffusion models [215], as demonstrated in a very recent work [216], offer potential for learning disentangled representations by modeling the evolution of probability distributions over time. These models could be further explored in the context of disentanglement to separate identity-related features from other factors. This area of research remains relatively unexplored and presents an opportunity to develop more effective techniques for privacy-preserving speaker verification while maintaining

identity integrity.

- **Privacy Protection against Informed Attacks:** The findings from the simulated informed attacks in Chapter 5 highlight the need to enhance privacy measures against attackers with knowledge of the protection methods used. Future efforts should concentrate on strengthening the concealment of sensitive attributes like gender. To address this concern, we can incorporate additional layers of obfuscation to hide sensitive attributes effectively.

- **Development of Bias Mitigation Techniques:** The evaluation of fairness in ASV systems, as presented in Chapters 5 and 6, has revealed disparities in outcomes among demographic groups, underscoring the imperative for implementing *fairness by design*. Building upon this evaluation, the implementation of bias mitigation techniques tailored for ASV systems becomes crucial. The literature review in Section 2.2.4 highlights a handful of proposed bias mitigation methods in the ASV domain. With insights gained from Chapter 6 regarding suitable fairness evaluation metrics for biometric recognition systems, including ASV, there emerges a clearer framework for comparing system fairness and selecting appropriate metrics. Moving forward, the development of specific bias mitigation strategies for ASV systems is essential to address biases that may result in inequitable outcomes across different demographic groups.

- **Investigation into Explainability Methods:** The introduction of the Fairness Activation Discrepancy (FAD) metric in Chapter 5, aimed at studying fairness across network layers, has raised questions about the underlying reasons for bias within ASV systems. This prompts the need to explore explainability methods to elucidate the origins of bias within these models. Understanding the contributing factors to bias is essential for developing effective mitigation strategies. Moreover, such efforts align with the GDPR principle of transparency, which mandates providing meaningful information about the logic behind automated decision-making processes to data subjects [217].

Bibliography

[1] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[2] Y. Chen, S. Zheng, H. Wang, L. Cheng, Q. Chen, and J. Qi, "An Enhanced Res2Net with Local and Global Feature Fusion for Speaker Verification," in *Proc. INTERSPEECH 2023*, 2023, pp. 2228–2232.

[3] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, "Cam++: A fast and efficient network for speaker verification using context-aware masking," 2023.

[4] C. Antal-Vaida, "A review of artificial intelligence and machine learning adoption in banks, during the covid-19 outbreak," in *Proceedings of the International Conference on Business Excellence*, vol. 16, no. 1, 2022, pp. 1316–1328.

[5] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[6] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[7] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[9] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of audio deepfake detection," in *Proc. Speaker Odyssey*, 2020, pp. 1–5.

[10] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. INTERSPEECH 2015*, 2015, pp. 3214–3218.

[11] S. Ioffe, "Probabilistic linear discriminant analysis," in *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part IV 9*. Springer, 2006, pp. 531–542.

[12] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[13] A. Higgins, L. Bahler, and J. Porter, "Speaker verification using randomized phrase prompting," *Digital signal processing*, vol. 1, no. 2, pp. 89–106, 1991.

[14] A. E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F. K. Soong, "The use of cohort normalized scores for speaker verification." in *ICSLP*, vol. 92, 1992, pp. 599–602.

[15] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, pp. 1–22, 2004.

[16] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[17] L. Li, R. Wang, G. Wang, C. Wang, and T. F. Zheng, "Decision making based on cohort scores for speaker verification," in *2016 Asia-Pacific Sig-*

nal and Information Processing Association Annual Summit and Conference (APSIPA). IEEE, 2016, pp. 1–4.

[18] M. J. Alam, G. Bhattacharya, and P. Kenny, "Speaker verification in mismatched conditions with frustratingly easy domain adaptation." in *Odyssey*, vol. 2018, 2018, pp. 176–180.

[19] P. Regulation, "Regulation (eu) 2016/679 of the european parliament and of the council," *Regulation (eu)*, vol. 679, p. 2016, 2016.

[20] T. Madiega, "Artificial intelligence act," *European Parliament: European Parliamentary Research Service*, 2021.

[21] K. Nandakumar and A. K. Jain, *Soft Biometrics*. Boston, MA: Springer US, 2009, pp. 1235–1239. [Online]. Available: https://doi.org/10.1007/978-0-387-73003-5_225

[22] ——, *Soft Biometrics*. Boston, MA: Springer US, 2009, pp. 1235–1239. [Online]. Available: https://doi.org/10.1007/978-0-387-73003-5_225

[23] S. R. Zaman, D. Sadekeen, M. A. Alfaz, and R. Shahriyar, "One source to detect them all: Gender, age, and emotion detection from voice," in *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2021, pp. 338–343.

[24] V. Mikhailava, M. Lesnichaia, N. Bogach, I. Lezhenin, J. Blake, and E. Pyshkin, "Language accent detection with cnn using sparse data from a crowd-sourced speech archive," *Mathematics*, vol. 10, no. 16, p. 2913, 2022.

[25] F. A. Shaqra, R. Duwairi, and M. Al-Ayyoub, "Recognizing emotion from speech based on age and gender using hierarchical models," *Procedia Computer Science*, vol. 151, pp. 37–44, 2019.

[26] G. Solana-Lavalle and R. Rosas-Romero, "Analysis of voice as an assisting tool for detection of parkinson's disease and its subsequent clinical interpretation," *Biomedical Signal Processing and Control*, vol. 66, p. 102415, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421000124

[27] L. M. Mazaira-Fernandez, A. Álvarez-Marquina, and P. Gómez-Vilda, "Improving speaker recognition by biometric voice deconstruction," *Frontiers in bioengineering and biotechnology*, vol. 3, p. 126, 2015.

[28] M. Najafian and M. Russell, "Automatic accent identification as an analytical tool for accent robust automatic speech recognition," *Speech Communication*, vol. 122, pp. 44–55, 2020.

[29] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, "Speech2face: Learning the face behind a voice," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7539–7548.

[30] L. Lima, V. Furtado, E. Furtado, and V. Almeida, "Empirical analysis of bias in voice-based personal assistants," in *Companion Proceedings of the 2019 World Wide Web Conference*, 2019, pp. 533–538.

[31] W. T. Hutiri and A. Y. Ding, "Bias in automated speaker recognition," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 230–247.

[32] M. Westerlund, "The emergence of deepfake technology: A review," *Technology innovation management review*, vol. 9, no. 11, 2019.

[33] P. Patrick, G. Aversano, R. Blouet, M. Charbit, and G. Chollet, "Voice forgery using alisp: indexation in a client memory," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. I–17.

[34] T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi, "On the security of hmm-based speaker verification systems against imposture using synthetic speech." in *Eurospeech*, 1999, pp. 1223–1226.

[35] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[36] I. Serna, A. Pena, A. Morales, and J. Fierrez, "Insidebias: Measuring bias in deep networks and application to face gender biometrics," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 3720–3727.

[37] A. C. Yao, "Protocols for secure computations," in *23rd annual symposium on foundations of computer science (sfcs 1982)*. IEEE, 1982, pp. 160–164.

[38] S. Goldwasser, "How to play any mental game, or a completeness theorem for protocols with an honest majority," *Proc. the Nineteenth Annual ACM STOC'87*, pp. 218–229, 1987.

[39] Y. Lindell, "Secure multiparty computation (mpc)," *Cryptology ePrint Archive*, 2020.

[40] A. Yao, "How to generate and exchange secrets (extended abstract)," in *FOCS*, 1986.

[41] D. Beaver, "Efficient multiparty protocols using circuit randomization," in *Advances in Cryptology—CRYPTO'91: Proceedings 11*. Springer, 1992, pp. 420–432.

[42] D. Evans, V. Kolesnikov, M. Rosulek *et al.*, "A pragmatic introduction to secure multi-party computation," *Foundations and Trends® in Privacy and Security*, vol. 2, no. 2-3, pp. 70–246, 2018.

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

[44] M. Aliasgari, M. Blanton, and F. Bayatbabolghani, "Secure computation of hidden markov models and secure floating-point arithmetic in the malicious model," *International Journal of Information Security*, vol. 16, pp. 577–601, 2017.

[45] A. Treiber, A. Nautsch, J. Kolberg, T. Schneider, and C. Busch, "Privacy-preserving plda speaker verification using outsourced secure computation," *Speech Communication*, vol. 114, pp. 60–71, 2019.

[46] A. Nautsch, J. Patino, A. Treiber, T. Stafylakis, P. Mizera, M. Todisco, T. Schneider, and N. Evans, "Privacy-Preserving Speaker Recognition with Cohort Score Normalisation," in *Proc. Interspeech 2019*, 2019, pp. 2868–2872.

[47] F. Teixeira, A. Abad, B. Raj, and I. Trancoso, "Towards End-to-End Private Automatic Speaker Recognition," in *Proc. Interspeech 2022*, 2022, pp. 2798–2802.

[48] R. L. Rivest, L. Adleman, M. L. Dertouzos *et al.*, "On data banks and privacy homomorphisms," *Foundations of secure computation*, vol. 4, no. 11, pp. 169–180, 1978.

[49] C. Gentry, *A fully homomorphic encryption scheme*. Stanford university, 2009.

[50] ——, "Fully homomorphic encryption using ideal lattices," in *Proceedings of the forty-first annual ACM symposium on Theory of computing*, 2009, pp. 169–178.

[51] C. Fontaine and F. Galand, "A survey of homomorphic encryption for non-specialists," *EURASIP Journal on Information Security*, vol. 2007, pp. 1–10, 2007.

[52] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," *Communications of the ACM*, vol. 21, no. 2, pp. 120–126, 1978.

[53] T. ElGamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," *IEEE transactions on information theory*, vol. 31, no. 4, pp. 469–472, 1985.

[54] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *International conference on the theory and applications of cryptographic techniques*. Springer, 1999, pp. 223–238.

[55] A. Acar, H. Aksu, A. S. Uluagac, and M. Conti, "A survey on homomorphic encryption schemes: Theory and implementation," *ACM Computing Surveys (Csur)*, vol. 51, no. 4, pp. 1–35, 2018.

[56] D. Boneh, E.-J. Goh, and K. Nissim, "Evaluating 2-dnf formulas on ciphertexts," in *Theory of Cryptography: Second Theory of Cryptography Conference, TCC 2005, Cambridge, MA, USA, February 10-12, 2005. Proceedings 2*. Springer, 2005, pp. 325–341.

[57] M. R. Albrecht, P. Farshim, J.-C. Faugere, and L. Perret, "Polly cracker, revisited," in *International Conference on the Theory and Application of Cryptology and Information Security*. Springer, 2011, pp. 179–196.

[58] C. Marcolla, V. Sucasas, M. Manzano, R. Bassoli, F. H. Fitzek, and N. Aaraj, "Survey on fully homomorphic encryption, theory, and applications," *Proceedings of the IEEE*, vol. 110, no. 10, pp. 1572–1609, 2022.

[59] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *Proceedings of the forty-first annual ACM symposium on Theory of computing*, 2009, pp. 169–178.

[60] Z. Brakerski, C. Gentry, and V. Vaikuntanathan, "(leveled) fully homomorphic encryption without bootstrapping," *ACM Transactions on Computation Theory (TOCT)*, vol. 6, no. 3, pp. 1–36, 2014.

[61] J. Fan and F. Vercauteren, "Somewhat practical fully homomorphic encryption," *Cryptology ePrint Archive*, 2012.

[62] Z. Brakerski, "Fully homomorphic encryption without modulus switching from classical gapsvp," in *Annual Cryptology Conference*. Springer, 2012, pp. 868–886.

[63] J. H. Cheon, A. Kim, M. Kim, and Y. Song, "Homomorphic encryption for arithmetic of approximate numbers," in *Advances in Cryptology–ASIACRYPT 2017: 23rd International Conference on the Theory and Applications of Cryptology and Information Security, Hong Kong, China, December 3-7, 2017, Proceedings, Part I 23*. Springer, 2017, pp. 409–437.

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

[65] ——, "Privacy-preserving speaker verification and identification using gaussian mixture models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 397–406, 2012.

[66] M. Pathak, J. Portelo, B. Raj, and I. Trancoso, "Privacy-preserving speaker authentication," in *Information Security: 15th International Conference, ISC 2012, Passau, Germany, September 19-21, 2012. Proceedings 15*. Springer, 2012, pp. 1–22.

[67] A. Nautsch, S. Isadskiy, J. Kolberg, M. Gomez-Barrero, and C. Busch, "Homomorphic Encryption for Speaker Recognition: Protection of Biometric Templates and Vendor Model Parameters ," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2018)*, 2018, pp. 16–23.

[68] Y. Rahulamathavan, "Privacy-preserving similarity calculation of speaker features using fully homomorphic encryption," *arXiv preprint arXiv:2202.07994*, 2022.

[69] C. Dwork, "Differential privacy," in *International colloquium on automata, languages, and programming*. Springer, 2006, pp. 1–12.

[70] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer, 2006, pp. 265–284.

[71] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy." *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3-4, pp. 211–407, 2014.

[72] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" *SIAM Journal on Computing*, vol. 40, no. 3, pp. 793–826, 2011.

[73] K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth sensitivity and sampling in private data analysis," in *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, 2007, pp. 75–84.

[74] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer, 2006, pp. 265–284.

[75] K. Zhu, P. Van Hentenryck, and F. Fioretto, "Bias and variance of post-processing in differential privacy," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 11 177–11 184.

[76] J. M. Abowd, "The us census bureau adopts differential privacy," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2867–2867.

[77] A. Ziller, D. Usynin, R. Braren, M. Makowski, D. Rueckert, and G. Kaissis, "Medical imaging deep learning with differential privacy," *Scientific Reports*, vol. 11, no. 1, pp. 1–8, 2021.

[78] B. Jiang, J. Li, G. Yue, and H. Song, "Differential privacy for industrial internet of things: Opportunities, applications, and challenges," *IEEE Internet of Things Journal*, vol. 8, no. 13, pp. 10 430–10 451, 2021.

[79] M. Yang, T. Guo, T. Zhu, I. Tjuawinata, J. Zhao, and K.-Y. Lam, "Local differential privacy and its applications: A comprehensive survey," *Computer Standards & Interfaces*, p. 103827, 2023.

[80] M. A. P. Chamikara, P. Bertok, I. Khalil, D. Liu, and S. Camtepe, "Privacy preserving face recognition utilizing differential privacy," *Computers & Security*, vol. 97, p. 101951, 2020.

[81] A. S. Shamsabadi, B. M. L. Srivastava, A. Bellet, N. Vauquier, E. Vincent, M. Maouche, M. Tommasi, and N. Papernot, "Differentially private speaker anonymization," *arXiv preprint arXiv:2202.11823*, 2022.

[82] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *Twelfth annual conference of the international speech communication association*, 2011.

[83] M. Azraoui, M. Bahram, B. Bozdemir, S. Canard, E. Ciceri, O. Ermis, R. Masalha, M. Mosconi, M. Önen, M. Paindavoine *et al.*, "Sok: Cryptography for neural networks," in *IFIP International Summer School on Privacy and Identity Management*. Springer, 2019, pp. 63–81.

[84] B. Bozdemir, "Privacy-preserving machine learning techniques," Ph.D. dissertation, Sorbonne université, 2021.

[85] S. Wagh, S. Tople, F. Benhamouda, E. Kushilevitz, P. Mittal, and T. Rabin, "Falcon: Honest-majority maliciously secure framework for private deep learning," *Proceedings on Privacy Enhancing Technologies*, vol. 2021, pp. 188–208, 01 2021.

[86] S. Wagh, D. Gupta, and N. Chandran, "Securenn: 3-party secure computation for neural network training." *Proc. Priv. Enhancing Technol.*, vol. 2019, no. 3, pp. 26–49, 2019.

[87] B. Bozdemir, O. Ermis, and M. Önen, "ProteiNN: Privacy-preserving one-to-many Neural Network classifications," in *SECRYPT 2020, 17th International Joint Conference on Security and Cryptography*, Lieusaint (on line), France, Jul. 2020. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03151068

[88] A. Boudguiga., O. Stan., A. Fazzat., H. Labiod., and P. Clet., "Privacy preserving services for intelligent transportation systems with homomorphic encryption," in *Proceedings of the 7th International Conference on Information Systems Security and Privacy - Volume 1: ICISSP,*, INSTICC. SciTePress, 2021, pp. 684–693.

[89] C. Juvekar, V. Vaikuntanathan, and A. Chandrakasan, "GAZELLE: A low latency framework for secure neural network inference," in *27th USENIX Security Symposium (USENIX Security 18)*. Baltimore, MD: USENIX Association, Aug. 2018, pp. 1651–1669. [Online]. Available: https://www.usenix.org/conference/usenixsecurity18/presentation/juvekar

[90] P. Mishra, R. Lehmkuhl, A. Srinivasan, W. Zheng, and R. A. Popa, "Delphi: A cryptographic inference system for neural networks," in *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, 2020, pp. 27–30.

[91] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[92] X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu, "Disentangled representation learning," *arXiv preprint arXiv:2211.11695*, 2022.

[93] Y. Bengio, "Deep learning of representations: Looking forward," in *International conference on statistical language and speech processing*. Springer, 2013, pp. 1–37.

[94] J. Williams, "Learning disentangled speech representations," 2022.

[95] N. Hou, C. Xu, E. S. Chng, and H. Li, "Learning disentangled feature representations for speech enhancement via adversarial training," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 666–670.

[96] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International conference on machine learning*. PMLR, 2018, pp. 5180–5189.

[97] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen *et al.*, "Hierarchical generative modeling for controllable speech synthesis," *arXiv preprint arXiv:1810.07217*, 2018.

[98] W.-N. Hsu, Y. Zhang, R. J. Weiss, Y.-A. Chung, Y. Wang, Y. Wu, and J. Glass, "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5901–5905.

[99] Y. Kwon, S.-W. Chung, and H.-G. Kang, "Intra-Class Variation Reduction of Speaker Representation in Disentanglement Framework," in *Proc. Interspeech 2020*, 2020, pp. 3231–3235.

[100] M. Sang, W. Xia, and J. H. Hansen, "Deaan: Disentangled embedding and adversarial adaptation network for robust speaker representation learning," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6169–6173.

[101] T. Liu, K. A. Lee, Q. Wang, and H. Li, "Disentangling voice and content with self-supervision for speaker recognition," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[102] S. H. Mun, M. H. Han, M. Kim, D. Lee, and N. S. Kim, "Disentangled speaker representation learning via mutual information minimization," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 89–96.

[103] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[104] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[105] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International conference on learning representations*, 2016.

[106] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-vae," *arXiv preprint arXiv:1804.03599*, 2018.

[107] H. Kim and A. Mnih, "Disentangling by factorising," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2649–2658.

[108] A. Kumar, P. Sattigeri, and A. Balakrishnan, "Variational inference of disentangled latent concepts from unlabeled observations," *arXiv preprint arXiv:1711.00848*, 2017.

[109] R. T. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," *Advances in neural information processing systems*, vol. 31, 2018.

[110] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *Advances in neural information processing systems*, vol. 29, 2016.

[111] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1415–1424.

[112] Z. Lin, K. Thekumparampil, G. Fanti, and S. Oh, "Infogan-cr and modelcentrality: Self-supervised model training and selection for disentangling gans," in *international conference on machine learning*. PMLR, 2020, pp. 6127–6139.

[113] I. Jeon, W. Lee, M. Pyeon, and G. Kim, "Ib-gan: Disentangled representation learning with information bottleneck generative adversarial networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 7926–7934.

[114] X. Zhu, C. Xu, and D. Tao, "Where and what? examining interpretable disentangled representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5861–5870.

[115] K. Hashimoto, J. Yamagishi, and I. Echizen, "Privacy-preserving sound to degrade automatic speaker verification performance," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5500–5504.

[116] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, X.-Y. Li, Y. Wang, and Y. Deng, "Voicemask: Anonymize and sanitize voice input on mobile devices," *arXiv preprint arXiv:1711.11460*, 2017.

[117] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker Anonymisation Using the McAdams Coefficient," in *Proc. Interspeech 2021*, 2021, pp. 1099–1103.

[118] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker Anonymization Using X-vector and Neural Waveform Models," in *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, 2019, pp. 155–160.

[119] Y. Han, S. Li, Y. Cao, Q. Ma, and M. Yoshikawa, "Voice-indistinguishability: Protecting voiceprint in privacy-preserving speech data release," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[120] B. M. L. Srivastava, N. Tomashenko, X. Wang, E. Vincent, J. Yamagishi, M. Maouche, A. Bellet, and M. Tommasi, "Design Choices for X-Vector Based Speaker Anonymization," in *Proc. Interspeech 2020*, 2020, pp. 1713–1717.

[121] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion?" in *Proc. Interspeech 2019*, 2019, pp. 3700–3704.

[122] R. Aloufi, H. Haddadi, and D. Boyle, "Privacy-preserving voice analysis via disentangled representations," in *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, 2020, pp. 1–14.

[123] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Introducing the VoicePrivacy Initiative," in *Proc. Interspeech 2020*, 2020, pp. 1693–1697.

[124] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J.-F. Bonastre, "The voiceprivacy 2022 challenge evaluation plan," *arXiv preprint arXiv:2203.12468*, 2022.

[125] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien *et al.*, "The voiceprivacy 2020 challenge: Results and findings," *Computer Speech & Language*, vol. 74, p. 101362, 2022.

[126] S. Meyer, F. Lux, P. Denisov, J. Koch, P. Tilli, and N. T. Vu, "Speaker Anonymization with Phonetic Intermediate Representations," in *Proc. Interspeech 2022*, 2022, pp. 4925–4929.

[127] X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, "Language-Independent Speaker Anonymization Approach Using Self-Supervised Pre-Trained Models," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 279–286.

[128] P. Champion, D. Jouvet, and A. Larcher, "Are disentangled representations all you need to build speaker anonymization systems?" in *INTERSPEECH 2022*, incheon, South Korea, Sep. 2022. [Online]. Available: https://hal.science/hal-03753746

[129] M. Tran and M. Soleymani, "Privacy-preserving Representation Learning for Speech Understanding," in *Proc. INTERSPEECH 2023*, 2023, pp. 2858–2862.

[130] M. Costante, M. Matassoni, and A. Brutti, "Using seq2seq voice conversion with pre-trained representations for audio anonymization: experimental insights," in *2022 IEEE International Smart Cities Conference (ISC2)*. IEEE, 2022, pp. 1–7.

[131] A. Broukhim and Z. Novack, "Towards generalizable deep speech anonymization."

[132] D. Tang, S. Zhou, H. Jiang, H. Chen, and Y. Liu, "Gender-adversarial networks for face privacy preserving," *IEEE Internet of Things Journal*, vol. 9, no. 18, pp. 17 568–17 576, 2022.

[133] L. Benaroya, N. Obin, and A. Roebel, "Beyond voice identity conversion: Manipulating voice attributes by adversarial learning of structured disentangled representations," *arXiv preprint arXiv:2107.12346*, 2021.

[134] R. Aloufi, H. Haddadi, and D. Boyle, "Privacy preserving speech analysis using emotion filtering at the edge: poster abstract," in *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, ser. SenSys '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 426–427. [Online]. Available: https://doi.org/10.1145/3356250.3361947

[135] V. Mirjalili, S. Raschka, and A. Ross, "Gender privacy: An ensemble of semi adversarial networks for confounding arbitrary gender classifiers," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2018, pp. 1–10.

[136] V. Mirjalili, S. Raschka, A. Namboodiri, and A. Ross, "Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images," in *2018 International Conference on Biometrics (ICB)*. IEEE, 2018, pp. 82–89.

[137] V. Mirjalili, S. Raschka, and A. Ross, "Flowsan: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers," *IEEE Access*, vol. 7, pp. 99 735–99 745, 2019.

[138] P. Melzi, H. O. Shahreza, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, J. Fierrez, S. Marcel, and C. Busch, "Multi-ive: Privacy enhancement of multiple soft-biometrics in face embeddings," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 323–331.

[139] P.-G. Noé, M. Mohammadamini, D. Matrouf, T. Parcollet, A. Nautsch, and J.-F. Bonastre, "Adversarial Disentanglement of Speaker Representation for Attribute-Driven Privacy Preservation," in *Proc. Interspeech 2021*, 2021, pp. 1902–1906.

[140] B. Bortolato, M. Ivanovska, P. Rot, J. Križaj, P. Terhörst, N. Damer, P. Peer, and V. Štruc, "Learning privacy-enhancing face representations through feature disentanglement," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 495–502.

[141] P. Terhörst, N. Damer, F. Kirchbuchner, and A. Kuijper, "Suppressing gender and age in face templates using incremental variable elimination," in *2019 International Conference on Biometrics (ICB)*. IEEE, 2019, pp. 1–8.

[142] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, "Biometrics recognition using deep learning: A survey," *Artificial Intelligence Review*, vol. 56, no. 8, pp. 8647–8695, 2023.

[143] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Demographic bias in biometrics: A survey on an emerging challenge," *IEEE Transactions on Technology and Society*, vol. 1, no. 2, pp. 89–103, 2020.

[144] P. Grother, M. Ngan, and K. Hanaoka, *Face recognition vendor test (fvrt): Part 3, demographic effects*. National Institute of Standards and Technology Gaithersburg, MD, 2019.

[145] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[146] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," in *Ethics of data and analytics*. Auerbach Publications, 2022, pp. 254–264.

[147] P. Garg, J. Villasenor, and V. Foggo, "Fairness metrics: A comparative analysis," in *2020 IEEE international conference on big data (Big Data)*. IEEE, 2020, pp. 3662–3666.

[148] J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell, "Fairness under unawareness: Assessing disparity when protected class is unobserved," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 339–348.

[149] N. Kallus, X. Mao, and A. Zhou, "Assessing algorithmic fairness with unobserved protected class using data combination," *Management Science*, vol. 68, no. 3, pp. 1959–1981, 2022.

[150] L. Zhang, Y. Wu, and X. Wu, "Situation testing-based discrimination discovery: A causal inference approach." in *IJCAI*, vol. 16, 2016, pp. 2718–2724.

[151] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[152] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[153] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[154] J. R. Loftus, C. Russell, M. J. Kusner, and R. Silva, "Causal reasoning for algorithmic fairness," *arXiv preprint arXiv:1805.05859*, 2018.

[155] A. Khademi, S. Lee, D. Foley, and V. Honavar, "Fairness in algorithmic decision making: An excursion through the lens of causality," in *The World Wide Web Conference*, 2019, pp. 2907–2914.

[156] Y. Wu, L. Zhang, and X. Wu, "Counterfactual fairness: Unidentification, bound and algorithm," in *Proceedings of the twenty-eighth international joint conference on Artificial Intelligence*, 2019.

[157] Y. Wu, L. Zhang, X. Wu, and H. Tong, "Pc-fairness: A unified framework for measuring causality-based fairness," *Advances in neural information processing systems*, vol. 32, 2019.

[158] X. Wang, Y. Zhang, and R. Zhu, "A brief review on algorithmic fairness," *Management System Engineering*, vol. 1, no. 1, p. 7, 2022.

[159] M. H. Teodorescu, L. Morse, Y. Awwad, and G. C. Kane, "Failures of fairness in automation require a deeper understanding of human-ml augmentation." *MIS quarterly*, vol. 45, no. 3, 2021.

[160] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt, "Delayed impact of fair machine learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3150–3158.

[161] B. Green and L. Hu, "The myth in the methodology: Towards a recontextualization of fairness in machine learning," in *Proceedings of the machine learning: the debates workshop*, 2018.

[162] J. J. Howard, Y. B. Sirotin, and A. R. Vemury, "The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance," in *2019 ieee 10th international conference on biometrics theory, applications and systems (btas)*. IEEE, 2019, pp. 1–8.

[163] T. de Freitas Pereira and S. Marcel, "Fairness in biometrics: A figure of merit to assess biometric verification systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 1, pp. 19–29, 2022.

[164] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC DIS 19795-10. Information Technology – Biometric Performance Testing and Reporting – Part 10: Quantifying biometric system performance variation across demographic groups*, International Organization for Standardization, 2023.

[165] J. Howard, E. Laird, Y. Sirotin, R. Rubin, J. Tipton, and A. Vemury, "Evaluating proposed fairness models for face recognition algorithms," in *Proc. Intl. Conf. on Pattern Recognition*, 2022.

[166] K. Kotwal and S. Marcel, "Fairness index measures to evaluate bias in biometric recognition," in *International Conference on Pattern Recognition*. Springer, 2022, pp. 479–493.

[167] G. Fenu, H. Lafhouli, and M. Marras, "Exploring algorithmic fairness in deep speaker verification," in *Computational Science and Its Applications–ICCSA 2020: 20th International Conference, Cagliari, Italy, July 1–4, 2020, Proceedings, Part IV 20*. Springer, 2020, pp. 77–93.

[168] G. Fenu, G. Medda, M. Marras, and G. Meloni, "Improving fairness in speaker recognition," in *Proceedings of the 2020 European Symposium on Software Engineering*, 2020, pp. 129–136.

[169] W. T. Hutiri and A. Y. Ding, "Bias in automated speaker recognition," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 230–247.

[170] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2021, pp. 1–9.

[171] A. Nagrani, J. S. Chung, J. Huh, A. Brown, E. Coto, W. Xie, M. McLaren, D. A. Reynolds, and A. Zisserman, "Voxsrc 2020: The second voxceleb speaker recognition challenge," *arXiv preprint arXiv:2012.06867*, 2020.

[172] G. Fenu, M. Marras, G. Medda, G. Meloni *et al.*, "Fair voice biometrics: Impact of demographic imbalance on group fairness in speaker recognition," in *Interspeech*. International Speech Communication Association, 2021, pp. 1892–1896.

[173] W. Toussaint and A. Y. Ding, "Sveva fair: A framework for evaluating fairness in speaker verification," *arXiv preprint arXiv:2107.12049*, 2021.

[174] O. Sadjadi, C. Greenberg, E. Singer, L. Mason, and D. Reynolds, "Nist 2021 speaker recognition evaluation plan," 2021.

[175] R. Peri, K. Somandepalli, and S. Narayanan, "A study of bias mitigation strategies for speaker recognition," *Computer Speech & Language*, vol. 79, p. 101481, 2023.

[176] M. Estevez and L. Ferrer, "Study on the fairness of speaker verification systems across accent and gender groups," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[177] H. Shen, Y. Yang, G. Sun, R. Langman, E. Han, J. Droppo, and A. Stolcke, "Improving fairness in speaker verification via group-adapted fusion network," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7077–7081.

[178] M. Jin, C. J.-T. Ju, Z. Chen, Y.-C. Liu, J. Droppo, and A. Stolcke, "Adversarial reweighting for speaker verification fairness," *arXiv preprint arXiv:2207.07776*, 2022.

[179] W. T. Hutiri, L. Gorce, and A. Y. Ding, "Design guidelines for inclusive speaker verification evaluation datasets," *arXiv preprint arXiv:2204.02281*, 2022.

[180] "ISO/IEC 30107-1: Information Technology - Biometric Presentation Attack Detection - Part 1: Framework," International Organization for Standardization, Geneva, Switzerland, 2023, standard.

[181] "ISO/IEC 19795-1:2021: Information Technology - Biometric Performance Testing and Reporting - Part 1: Principles and Framework," International Organization for Standardization, Geneva, Switzerland, 2021, standard.

[182] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *speech communication*, vol. 66, pp. 130–153, 2015.

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

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

[185] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "Asvspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.

[186] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch *et al.*, "Asvspoof 2021:

Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[187] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *Proc. ICASSP*, 2021.

[188] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification ," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2018)*, 2018, pp. 312–319.

[189] T. Kinnunen, H. Delgado, N. Evans, K. A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, M. Sahidullah, J. Yamagishi *et al.*, "Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2195–2210, 2020.

[190] M. Todisco, H. Delgado, and N. Evans, "Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[191] ——, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Speaker Odyssey Workshop*, vol. 25, Bilbao, Spain, 2016, pp. 249–252.

[192] ——, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech and Language*, vol. 45, pp. 516–535, 2017.

[193] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2087–2091.

[194] H. Tak, J. Patino, A. Nautsch *et al.*, "Spoofing Attack Detection using the Non-linear Fusion of Sub-band Classifiers," in *Proc. INTERSPEECH*, 2020, pp. 1106–1110.

[195] T. Ryffel, A. Trask, M. Dahl, B. Wagner, J. Mancuso, D. Rueckert, and J. Passerat-Palmbach, "A generic framework for privacy preserving deep learning," *arXiv preprint arXiv:1811.04017*, 2018.

[196] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[197] I. Damgård, V. Pastro, N. Smart, and S. Zakarias, "Multiparty computation from somewhat homomorphic encryption," in *Annual Cryptology Conference*. Springer, 2012, pp. 643–662.

[198] I. Damgård, M. Keller, E. Larraia, V. Pastro, P. Scholl, and N. P. Smart, "Practical covertly secure mpc for dishonest majority–or: breaking the spdz limits," in *Computer Security–ESORICS 2013: 18th European Symposium on Research in Computer Security, Egham, UK, September 9-13, 2013. Proceedings 18*. Springer, 2013, pp. 1–18.

[199] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 308–318. [Online]. Available: https://doi.org/10.1145/2976749.2978318

[200] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 656–672.

[201] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[202] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[203] R. K. Das, R. Tao, and H. Li, "Hlt-nus submission for 2020 nist conversational telephone speech sre," *arXiv preprint arXiv:2111.06671*, 2021.

[204] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[205] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=rkE3y85ee

[206] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 1652–1656.

[207] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016. [Online]. Available: http://jmlr.org/papers/v17/15-239.html

[208] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, W. Ammar, A. Louis, and N. Mostafazadeh, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 48–53. [Online]. Available: https://aclanthology.org/N19-4009

[209] J. J. Howard, E. J. Laird, R. E. Rubin, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Evaluating proposed fairness models for face recognition algorithms," in *International Conference on Pattern Recognition*. Springer, 2022, pp. 431–447.

[210] G. Deltas, "The small-sample bias of the gini coefficient: results and implications for empirical research," *Review of economics and statistics*, vol. 85, no. 1, pp. 226–234, 2003.

[211] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," in *Proc. Interspeech*, 2020.

[212] W. Cai, J. Chen, and M. Li, "Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2018)*, 2018, pp. 74–81.

[213] Y. Kwon, H.-S. Heo, B.-J. Lee, and J. S. Chung, "The ins and outs of speaker recognition: lessons from voxsrc 2020," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5809–5813.

[214] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Proc. Interspeech 2018*, 2018, pp. 2252–2256.

[215] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[216] T. Yang, C. Lan, Y. Lu *et al.*, "Diffusion model with cross attention as an inductive bias for disentanglement," *arXiv preprint arXiv:2402.09712*, 2024.

[217] A. D. Selbst and J. Powles, "Meaningful information and the right to explanation," *International Data Privacy Law*, vol. 7, no. 4, pp. 233–242, 12 2017. [Online]. Available: https://doi.org/10.1093/idpl/ipx022