Differentially Private Adversarial Auto-Encoder to Protect Gender in Voice Biometrics

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

Oubaïda Chouchane
EURECOM
Sophia Antipolis, France
oubaida.chouchane@eurecom.fr

Michele Panariello
EURECOM
Sophia Antipolis, France
michele.panariello@eurecom.fr

Oualid Zari
oualid.zari@eurecom.fr
EURECOM
Sophia Antipolis, France

Ismet Kerenciler
EURECOM
Sophia Antipolis, France
Ismet.Kerenciler@eurecom.fr

Imen Chihaoui
EURECOM
Sophia Antipolis, France
Imen.Chihaoui@eurecom.fr

Massimiliano Todisco
EURECOM
Sophia Antipolis, France
massimiliano.todisco@eurecom.fr

Melek Önen
EURECOM
Sophia Antipolis, France
melek.onen@eurecom.fr

ABSTRACT

Over the last decade, the use of Automatic Speaker Verification (ASV) systems has become increasingly widespread in response to the growing need for secure and efficient identity verification methods. The voice data encompasses a wealth of personal information, which includes but is not limited to gender, age, health condition, stress levels, and geographical and socio-cultural origins. These attributes, known as soft biometrics, are private and the user may wish to keep them confidential. However, with the advancement of machine learning algorithms, soft biometrics can be inferred automatically, creating the potential for unauthorized use. As such, it is crucial to ensure the protection of these personal data that are inherent within the voice while retaining the utility of identity recognition. In this paper, we present an adversarial Auto-Encoder–based approach to hide gender-related information in speaker embeddings, while preserving their effectiveness for speaker verification. We use an adversarial procedure against a gender classifier and incorporate a layer based on the Laplace mechanism into the Auto-Encoder architecture. This layer adds Laplace noise for more robust gender concealment and ensures differential privacy guarantees during inference for the output speaker embeddings. Experiments conducted on the VoxCeleb dataset demonstrate that speaker verification tasks can be effectively carried out while concealing speaker gender and ensuring differential privacy guarantees; moreover, the intensity of the Laplace noise can be tuned to select the desired trade-off between privacy and utility.

IH&MMSec '23, June 28–30, 2023, Chicago, IL, USA.

ACM ISBN 979-8-4007-0054-5/23/06…$15.00
https://doi.org/10.1145/3577163.3595102

CCS CONCEPTS

• **Security and privacy → Biometrics**; **Privacy protections**; • **Computing methodologies → Machine learning**.

KEYWORDS

speaker verification; gender recognition; privacy preservation; differential privacy

ACM Reference Format:
Oubaïda Chouchane, Michele Panariello, Oualid Zari, Ismet Kerenciler, Imen Chihaoui, Massimiliano Todisco, and Melek Önen . 2023. Differentially Private Adversarial Auto-Encoder to Protect Gender in Voice Biometrics. In *Proceedings of the 46th International ACM IH&MMSec Conference on Research and Development in Information Retrieval (IH&MMSec '23), June 28–30, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3577163.3595102

1 INTRODUCTION

Voice is a unique biometric trait that is widely recognized for its capability to efficiently and securely identify individuals [26]. The use of voice as a biometric modality has been deployed in Automatic Speaker Verification (ASV) systems that have been incorporated into a range of applications like personal database access, credit card authorization, voice banking, and funds transfer. In speaker verification, also known as speaker authentication, a user claims their identity, and the system evaluates the truthfulness of that claim by comparing the speaker's biometric characteristics with the stored representation of the claimed identity. The system seeks to establish a match between the speaker's features and the claimed identity that surpasses a specified threshold. In instances where a match is not found, the speaker is rejected. Moreover, the voice does not only contain unique identity information but also physiological or psychological aspects like age, gender, emotions, accent, ethnicity, personality, and health condition, referred to as soft biometrics [22], that can be detected automatically using machine learning systems [32]. The utilization of these soft biometric traits in conjunction with primary biometrics can provide additional information for the recognition process and improve recognition

accuracy [15]. Studies in [25] also show that speakers' short recordings can be used to reconstruct their average-looking facial images that embody their physical characteristics such as age, gender, and ethnicity. However, despite their potential use for legitimate processing purposes, soft biometrics are susceptible to malicious utilization. This can occur through unauthorized data processing that puts individuals at risk of privacy concerns such as discrimination, invasive advertising, extortion, and other forms of abuse. As a specific illustration, the finance sector has been shown to exhibit gender-based biases in loan provision [6]. This raises concerns regarding the potential existence of discriminatory lending practices that pose greater barriers to women than to men in the pursuit of starting a business enterprise [9]. Solutions based on cryptographic primitives [7, 31], while effective, produce completely garbled messages. Data obfuscation techniques, on the other hand, provide a more balanced approach to privacy preservation, protecting sensitive information without rendering the entire message content unrecognizable. Moreover, the voice is recognized as personal and sensitive and is therefore subject to protection under the General Data Protection Regulation (GDPR or Regulation 2016/679)[1] together with numerous other data protection legislation, worldwide. The GDPR considers gender as well as a form of personal data and imposes an obligation to safeguard its protection. In light of the increasing concerns surrounding privacy, there has been a growing effort to protect private information like soft biometrics. This effort has led to multiple research initiatives aimed at developing and implementing effective techniques for protecting the privacy of soft biometric attributes [20, 30]. Among these, techniques based on the differential privacy (DP) notion [12] have received significant attention. Differentially private solutions (also referred to as global or centralized DP) were proposed for more than a decade and regarded as a privacy protection tool for different areas [2, 34]. While global DP mechanisms consist of a trusted central party/data curator collecting the users' data, aggregating them, and further protecting the aggregated information by adding some calibrated noise before releasing it to the public, local DP (LDP) solutions [13] protect the input data immediately to prevent the data curator from discovering the real, individual data. The noise is derived from a DP mechanism (e.g. Laplace mechanism). In this paper, we aim to address the challenge of protecting gender information while preserving the efficiency of speaker verification. Our approach is based on adding a calibrated noise drawn from the Laplace distribution during the training of an Adversarial Auto-Encoder (AAE) architecture. The noise is injected into the latent space (i.e. the output of the encoder) in order to assure that the model is ϵ-differentially private and to enhance the capability of the adversary in obscuring gender information. The speaker makes use of the private AAE locally to conceal their gender prior to the dissemination of their speaker features for the purpose of authentication. Our experiments conducted on the VoxCeleb 1 and VoxCeleb 2 datasets demonstrate the feasibility of executing speaker verification tasks effectively while disrupting adversarial attempts of gender recognition. To the best of our knowledge, this is the first work that uses differentially private solutions to protect gender information while preserving identity in biometrics.

[1] https://gdpr-info.eu/

2 RELATED WORK

In recent years, there has been a proliferation of academic literature pertaining to the topic of soft biometrics protection in biometric recognition systems. A significant number of researchers have centered their efforts on developing technical solutions that are capable of preventing the extraction of soft biometric attributes and are either directly applied to the collected biometric data like face images and voice signals (i.e. at sample level) [3, 4, 17, 28] or to the extracted features (i.e. at feature level) [5, 16, 24, 29].

Mirjalili et al. [17] proposed a Semi-Adversarial Network (SAN) based on an adversarial Convolutional Auto-Encoder (CAE) in order to hide the gender information from face images while retaining the biometric matching utility. In a follow-up work [18], the same authors introduced an ensemble of SANs that are constituted of multiple auxiliary gender classifiers and face matches that generates diverse perturbations for an input face image. The idea behind this approach is that at least one of the perturbed images succeeds in fooling an arbitrary gender classifier. In [19], Mirjalili et al. also attempted to combine a variety of face perturbations in an effort to improve the generalization capability of SAN models. Despite the successful privacy preservation of gender attributes by the aforementioned techniques, their robustness to arbitrary classifiers is limited. In a more recent study, Tang et al. [28] presented an alternative gender adversarial network model that effectively masks gender attributes while preserving both image quality and matching performance. Besides, this model demonstrates the ability to generalize to previously unseen gender classifiers. Further work was proposed by Bortolato et al. [5] to leverage the privacy-preservation of face images on the template level also using the AE technique. The authors suggested an AE-based solution that effectively separates gender attribute information from identity, resulting in good generalization performance across a variety of datasets. Additionally, Terhöst et al. [29] introduced an Incremental Variable Eliminations (IVE) algorithm that trains a set of decision trees to determine the importance of the variables that are crucial for predicting sensitive attributes. These variables were then incrementally removed from the facial templates to suppress gender and age features while maintaining high face-matching performance. In [16] Melzi et. al. extended this approach to protect multiple soft biometrics (i.e. gender, age, and ethnicity) present in facial images. In speech-related literature, Aloufi et al. [3] built a Voice Conversion (VC) system that can conceal the emotional state of the users while maintaining speech recognition utility for voice-controlled IoT. The model is based on a Cycle-Generative Adversarial Network (GAN) architecture. Similarly, in [4], the authors introduced a neural VC architecture that can manipulate gender attributes present in the voice signal. This proposed VC architecture involves multiple Auto-Encoders that transform speech into independent linguistic and extra-linguistic representations. These representations are learned through an adversarial process and can be adjusted during VC.

On a template level, Noé et. al. [24] proposed an adversarial Auto-Encoder architecture that disentangles gender attributes from x-vector speaker embeddings [27]. The AE is combined with an external gender classifier that attempts to predict the attribute class from the encoded representations. The proposed solution succeeds in concealing gender-related information in the embedding

while maintaining good ASV performance. Nonetheless, our experimental findings indicate that using speaker embeddings other than x-vectors, such as those generated by the ECAPA-TDNN model [10], yields inconsistent performance, implying potential challenges in achieving generalization. We hypothesize that this may be attributed to the superior representational capabilities of ECAPA-TDNN embeddings, which have largely superseded x-vectors in recent speaker modeling.

3 GENDER CONCEALMENT

In this section, we present the building blocks of the proposed gender concealment technique. First, we describe the architecture of the AAE and highlight its limitations in the concealment task. Second, we briefly introduce local differential privacy, a concept that is instrumental in improving the gender concealment capabilities of the model. Lastly, we illustrate how to combine the AAE and LDP to obtain a more effective technique for suppressing gender information in speaker embeddings, with a tunable privacy-utility trade-off and sound theoretical guarantees.

3.1 Gender-Adversarial Auto-Encoder

Let $\mathbf{x}$ be an embedding representing a speaker identity. The goal of a Gender-Adversarial Auto-Encoder is to process $\mathbf{x}$ so as to produce a new embedding $\tilde{\mathbf{x}}$ that still encodes the identity of that same speaker, but is devoid of any information about their gender. In this section, we describe our implementation of this system, which mostly follows the one proposed in [24].

Given an input embedding $\mathbf{x} \in \mathbb{R}^d$, we create a compressed representation of it by means of $e_{\phi_1}(\mathbf{x}) = \mathbf{z} \in \mathbb{R}^l$, where $e_{\phi_1}(\cdot)$ is a feed-forward neural network parameterized by ϕ_1 and $l < d$. The disentanglement of gender-related information from $\mathbf{z}$ depends on an adversarial "discriminator" module $a_\theta(\cdot)$ (also a feed-forward neural network) that attempts to infer the gender of the speaker associated with $\mathbf{z}$. During training, we optimize θ to minimize the objective:

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

where $y \in \{0, 1\}$ is the ground-truth gender label (0 for male, 1 for female) and $a_\theta(\mathbf{z}) \in [0, 1]$ represents the predicted probability of $\mathbf{z}$ having been produced by a female speaker. The suppression of the gender-related information is performed by adversarially training the encoder to "fool" the discriminator, i.e. to make it so that it is not capable of accurately predicting the speaker's gender from $\mathbf{z}$. In practice, this is achieved by optimizing the same objective as (1), except that the probability predicted by the discriminator is inverted:

$$\mathcal{L}_{adv}(\mathbf{x}, y, \phi_1 \mid \theta) = -y \log(1 - a_\theta(\mathbf{z})) - (1-y)\log(a_\theta(\mathbf{z})) \quad (2)$$

A decoder feed-forward 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, e.g. speaker verification, 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.

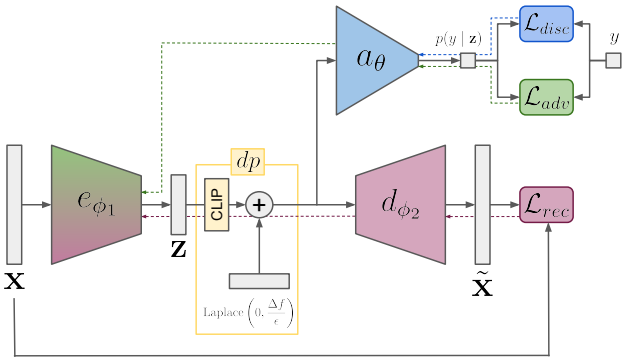

Figure 1: Illustration of the proposed system at training time. Solid and dashed arrows represent forward and backward propagation respectively. Modules are colored based on which gradient signal they are optimized by.

$$\mathcal{L}_{rec}(\mathbf{x}, \phi_1, \phi_2) = 1 - \cos\left(\mathbf{x}, d_{\phi_2}(\mathbf{z})\right) \quad (3)$$

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:

$$\begin{aligned} \phi &\leftarrow \nabla_\phi (\mathcal{L}_{adv} + \mathcal{L}_{rec}) \\ \theta &\leftarrow \nabla_\theta \mathcal{L}_{disc} \end{aligned} \quad (4)$$

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

$$\tilde{\mathbf{x}} = d_{\phi_2}\left(e_{\phi_1}(\mathbf{x})\right) \quad (5)$$

The privacy preservation capability of the Auto-Encoder is evaluated upon the ability of an attacker 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. 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 recognition = 98.45 (10^{-2}) and Equal Error Rate (EER) = 1.86% for ASV performance. In order to ensure that the predictions of the gender classifier are truly random, the AUC must be close to 50%. 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 which is well-studied for noise addition and calibration and also provides DP guarantees. The latent vectors $\mathbf{z}$ are locally differentially private thanks to the Laplace mechanism and subsequently, the reconstructed vectors are differentially private by the post-processing property of DP [33].

3.2 Local Differential Privacy

Local differential privacy plays a crucial role in protecting personal data like soft biometrics and assessing the privacy risks. In this section, we provide a brief description of the underlying concepts of local differential privacy and the Laplace mechanism.

Definition. Local differential privacy is a state-of-the-art privacy model and consists in protecting individual input data before its collection. LDP ensures privacy for each user locally (i.e. each individual record is protected rather than the entire dataset as a whole) by adding noise without the necessity of trusting a central authority. Formally, (ϵ)-local differential privacy is defined as follows.

Definition 3.1 ((Local Differential Privacy [13]). 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:

$$Pr[\mathcal{M}(x) \in S] \leq e^{\epsilon} \cdot Pr[\mathcal{M}(x') \in S] \tag{6}$$

where Pr denotes the probability and ϵ ($\epsilon > 0$) 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.

Sensitivity. The sensitivity [23], 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 sensitivity can be defined as shown in (7), where x and x' represent two adjacent records in a dataset $\mathcal{X}$ and $\|.\|$ denotes the ℓ_1 norm of a vector.

$$\Delta f = \max_{x,x' \in \mathcal{X}} \left\| f(x) - f(x') \right\|_1 \tag{7}$$

The 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. It defines the magnitude of the noise needed in order to meet the (ϵ)-LDP requirements.

Laplace Mechanism. The Laplace mechanism [11] is a widely adopted technique for achieving (ϵ)-LDP. The mechanism works by adding random noise, sampled from the Laplace distribution, 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. Formally, given a database $\mathcal{X}$ and a function $f : \mathcal{X} \rightarrow \mathbb{R}^d$ that maps the database to d real numbers, the Laplace mechanism is defined as:

$$\mathcal{M}(f(x), \epsilon) = f(x) + (n_1, n_2, ..., n_d). \tag{8}$$

where each $n_i \sim$ 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 [12]. In our work, we use the Laplace mechanism to perturb each component of the latent speaker embedding $\mathbf{z}$ with noise drawn from the Laplace distribution. This approach successfully conceals the speaker's gender while retaining the usefulness of the feature vectors for ASV tasks.

3.3 Gender-Adversarial Auto-Encoder with Laplace noise

We improve the gender concealment capability of the AAE by applying the Laplace mechanism to the latent space learned by the encoder. More specifically, during training, we pass the latent embedding $\mathbf{z}$ through a *Laplace* layer $dp\,(\cdot)$ that adds to its input a noisy vector $\mathbf{n} \sim$ Laplace $(0, \Delta f/\epsilon)$. Figure 1 graphically depicts the system.

As there is no prior bound on the ℓ_1-norm of the vector $\mathbf{z}$, we use the same clipping procedure described in [1]: it consists in scaling $\mathbf{z}$ by a coefficient $1/\max(1, \|\mathbf{z}\|_1/C)$, where C is the clipping threshold. 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. The purpose of the clipping is to ensure that the sensitivity between any pair of vectors $\mathbf{z}$ and $\mathbf{z'}$ is $\Delta f \leq 2C$. 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. Thus, the Laplace layer is defined as

$$dp\,(\mathbf{z}) = \frac{\mathbf{z}}{\max\left(1, \frac{\|\mathbf{z}\|_1}{C}\right)} + \mathbf{n} \tag{9}$$

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 3.1. Once the model has been trained, the adversarial module $a_\theta(\cdot)$ is removed. The value of ϵ can be chosen according to the desired balance between privacy protection and the utility of the produced embeddings.

The goal of adding Laplace noise into the system is twofold. At test time, its purpose is to provide privacy protection theoretical guarantees as previously described; however, at training time, it also serves as a regularizer for the adversarial module and the decoder. Indeed, our experiments show that applying the Laplace mechanism at training time only (i.e. removing the Laplace layer at test time) is sufficient to greatly enhance the gender concealment capabilities of the system.

To better explore the functional difference between the Laplace noise at test and at training time, we perform experiments by independently varying the value of ϵ during the training phase (ϵ_{tr}) and during the testing phase (ϵ_{ts}). Our results show that changing ϵ_{tr} is the most convenient way to roughly set the balance between the empirical capabilities of gender concealment and ASV performance; however, by definition, ϵ_{tr} does not provide full control over the DP budget of the embeddings at test time. Changing ϵ_{ts} then offers a flexible means of fine-tuning the privacy budget of the embeddings even once the model has been trained and deployed.

In Section 4, we show that both ϵ_{tr} and ϵ_{ts} are equally relevant in determining the behavior of the system.

Privacy Guarantees for the Gender-AAE. One of the main strengths of differential privacy lies in its property of post-processing, which ensures that the privacy guarantee offered by a DP mechanism remains unaltered regardless of the arbitrary computations performed on its output. More formally,

Definition 3.2 (Post-processing [11, 33]). Let $\mathcal{M}$ be an ϵ-differentially private mechanism and g be an arbitrary mapping from the set of

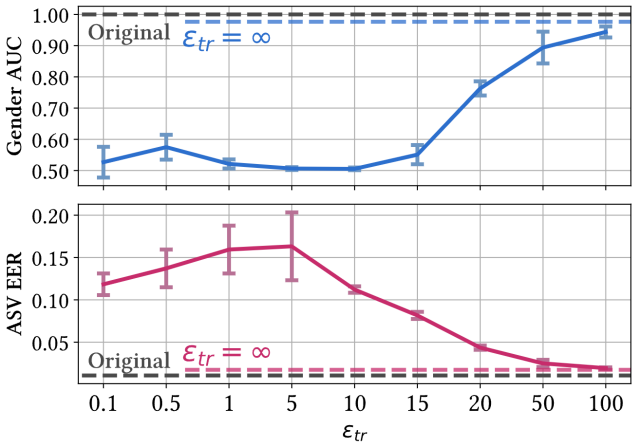

Figure 2: ASV EER and gender classification AUC achieved by the system for increasing values of ϵ_{tr}.

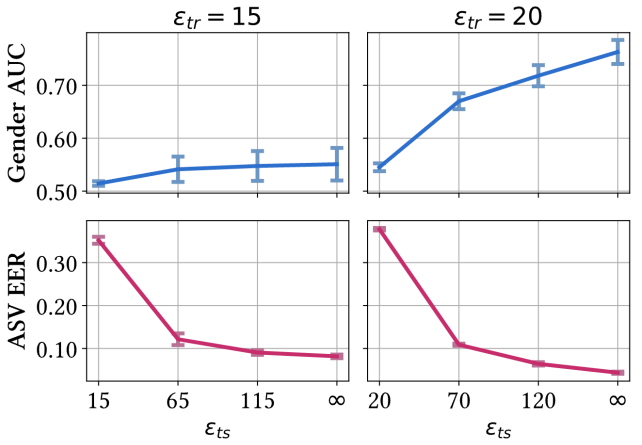

Figure 3: ASV EER and gender classification AUC achieved by the system for increasing values of ϵ_{ts}, for the cases of $\epsilon_{tr} = 15$ and $\epsilon_{tr} = 20$.

possible outputs to an arbitrary set. Then, $g \circ \mathcal{M}$ is ϵ-differentially private.

Similarly to the work in [14], we add noise to the latent space of the Auto-Encoder during the training, and use the same privacy proof, thanks to the post-processing property: $d_{\phi_2} \circ dp$ satisfies ϵ-DP, and so does the Auto-Encoder $d_{\phi_2} \circ dp \circ e_{\phi_1}$.

4 EXPERIMENTAL SETUP AND RESULTS

In this section, we discuss the experimental configurations and results. The feature extractor used to produce the speaker embeddings is the ECAPA-TDNN, whose output feature size is $d = 192$. The modules of the proposed encoder and decoder models are single-layer fully-connected neural networks and the gender classifiers (i.e. discriminator and external) are two-layer fully-connected neural networks. The encoder is followed by a ReLU activation and batch normalization, and the decoder is followed by a tanh activation function. We set the latent space to be of size $l = 64$. The adversarial classifier is composed of two fully-connected layers: the first one has 64 input units with a ReLU activation function, and the second one has 32 input units with a sigmoid activation function. An external gender classifier, used by an attacker to infer gender, is used to assess privacy protection and has the same architecture as the discriminator with 192 input units in the first layer and 100 input units in the second layer. The ASV assessment is done by first creating a model for each speaker; trial scores are then obtained by comparing trial embeddings with the respective speaker models by means of 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. The training dataset of the AAE is a subset of VoxCeleb2 [8] development partition (397032 segments per class). The testing is conducted using a subset of the VoxCeleb1 [21] test partition (2900 segments per class). The external sex classifier is trained using a subset of the VoxCeleb1 development partition (61616 segments per class). To select the clipping threshold C, we compute the median of the norm of all unclipped z vectors during the training, which is C=18.35.

We initially explore the behavior of the system by setting $\epsilon_{ts} = \infty$ (i.e. no DP protection) and for increasing values of ϵ_{tr}: Figure 2 shows the achieved ASV EER and gender classification AUC. We experimentally determine the noise scale and prioritize higher ϵ_{tr} resolution for the region with significant privacy/utility changes, while lower resolution suffices for regions with minor variations. As expected, privacy and utility scores inversely mirror one another. Specifically, $\epsilon_{tr} = 15$ seems to strike a satisfactory balance between the two, resulting in a 0.55 gender classification AUC while achieving an ASV EER of 8.1%. For comparison, the same gender classifier and ASV system obtain an AUC of nearly 1 and an EER of 1.1% on the original ECAPA embeddings, respectively.

We pick the model weights trained with $\epsilon_{tr} = 15$ and $\epsilon_{tr} = 20$ and experiment with values of $\epsilon_{ts} < \infty$ to add DP protection to the speaker embeddings. Setting $\epsilon_{ts} = \epsilon_{tr}$ further enhances the level of gender concealment: AUC scores drop from 0.55 to 0.50 (from 0.76 to 0.55 respectively) for $\epsilon_{tr} = \epsilon_{ts} = 15$ ($\epsilon_{tr} = \epsilon_{ts} = 20$ respectively). However, ASV EER degrades by around 20 percentage points in both scenarios. By increasing ϵ_{ts} by 20 units, it is possible to restore the ASV EER to around 10% (for both model versions) while achieving satisfactory AUC values of 0.55 and 0.68 for $\epsilon_{tr} = 15$ and $\epsilon_{tr} = 20$, respectively. In general, these results show the level of flexibility that the system can achieve even after training, all while providing DP guarantees over the produced embeddings.

Informal experiments run with $\epsilon_{tr} = \infty$ have resulted in rapid erasure of all meaningful information from the speaker embeddings even for high values of ϵ_{ts}: this is indicative of the relevance of including the Laplace noise during training for the DP protection to be applicable at test time.

5 CONCLUSIONS

We have presented an AE-based system to conceal gender-related information in speaker embeddings while retaining their utility for a speaker verification task. We perform the concealment by means of an adversarial game between an Auto-Encoder and an external gender classifier, and we improve upon previous work by introducing a Laplace-noise–addition layer within the architecture. The Laplace noise regularizes the training and allows for more

robust gender concealment, while also endowing the output speaker embedding with DP guarantees at inference time. The tuning of the ϵ parameter of the Laplace layer allows selecting the desired balance of privacy protection and utility, even after the training process has finished. Experimental results show that the proposed solution is effective in preserving gender privacy while maintaining utility for speaker verification tasks. Furthermore, the flexible trade-off between privacy and utility provided by our approach can be adapted to individual needs, making it a promising solution for privacy-preserving applications.

ACKNOWLEDGMENTS

This work is supported by the TReSPAsS-ETN project funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860813. It is also supported by the ANR-DFG RESPECT project.

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (Vienna, Austria) *(CCS '16)*. Association for Computing Machinery, New York, NY, USA, 308–318. https://doi.org/10.1145/2976749.2978318

[2] John M Abowd. 2018. The US Census Bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2867–2867.

[3] Ranya Aloufi, Hamed Haddadi, and David Boyle. 2019. Emotionless: Privacy-preserving speech analysis for voice assistants. *arXiv preprint arXiv:1908.03632* (2019).

[4] Laurent Benaroya, Nicolas Obin, and Axel Roebel. 2021. Beyond Voice Identity Conversion: Manipulating Voice Attributes by Adversarial Learning of Structured Disentangled Representations. *arXiv preprint arXiv:2107.12346* (2021).

[5] Blaž Bortolato, Marija Ivanovska, Peter Rot, Janez Križaj, Philipp Terhörst, Naser Damer, Peter Peer, and Vitomir Štruc. 2020. Learning privacy-enhancing face representations through feature disentanglement. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 495–502.

[6] J Michelle Brock and Ralph De Haas. 2021. EVIDENCE FROM BANKERS IN THE LAB. (2021).

[7] Oubaïda Chouchane, Baptiste Brossier, Jorge Esteban Gamboa Gamboa, Thomas Lardy, Hemlata Tak, Orhan Ermis, Madhu Kamble, Jose Patino, Nicholas Evans, Melek Önen, and Massimiliano Todisco. 2021. Privacy-Preserving Voice Anti-Spoofing Using Secure Multi-Party Computation. In *Interspeech 2021*. ISCA, Brno, Czech Republic, 856–860. https://doi.org/10.21437/Interspeech.2021-983

[8] J. S. Chung, A. Nagrani, and A. Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *INTERSPEECH*.

[9] Anastasia Cozarenco and Ariane Szafarz. 2018. Gender biases in bank lending: Lessons from microcredit in France. *Journal of Business Ethics* 147 (2018), 631–650.

[10] Nauman Dawalatabad, Mirco Ravanelli, François Grondin, Jenthe Thienpondt, Brecht Desplanques, and Hwidong Na. 2021. ECAPA-TDNN Embeddings for Speaker Diarization. In *Proc. Interspeech 2021*. 3560–3564. https://doi.org/10.21437/Interspeech.2021-941

[11] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. 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, 265–284.

[12] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* 9, 3-4 (2014), 211–407.

[13] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2011. What can we learn privately? *SIAM J. Comput.* 40, 3 (2011), 793–826.

[14] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2019. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 656–672.

[15] Luis Miguel Mazaira-Fernandez, Agustín Álvarez-Marquina, and Pedro Gómez-Vilda. 2015. Improving speaker recognition by biometric voice deconstruction. *Frontiers in bioengineering and biotechnology* 3 (2015), 126.

[16] Pietro Melzi, Hatef Otroshi Shahreza, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Sébastien Marcel, and Christoph Busch. 2023. Multi-IVE: Privacy Enhancement of Multiple Soft-Biometrics in Face Embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 323–331.

[17] Vahid Mirjalili, Sebastian Raschka, Anoop Namboodiri, and Arun Ross. 2018. Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images. In *2018 International Conference on Biometrics (ICB)*. IEEE, 82–89.

[18] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. 2018. 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, 1–10.

[19] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. 2019. Flowsan: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers. *IEEE Access* 7 (2019), 99735–99745.

[20] Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Ruben Tolosana. 2020. SensitiveNets: Learning agnostic representations with application to face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 6 (2020), 2158–2164.

[21] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612* (2017).

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

[23] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2007. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. 75–84.

[24] Paul-Gauthier Noé, Mohammad Mohammadamini, Driss Matrouf, Titouan Parcollet, Andreas Nautsch, and Jean-François Bonastre. 2020. Adversarial disentanglement of speaker representation for attribute-driven privacy preservation. *arXiv preprint arXiv:2012.04454* (2020).

[25] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T Freeman, Michael Rubinstein, and Wojciech Matusik. 2019. Speech2face: Learning the face behind a voice. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7539–7548.

[26] Mufan Sang, Yong Zhao, Gang Liu, John HL Hansen, and Jian Wu. 2023. Improving Transformer-based Networks With Locality For Automatic Speaker Verification. *arXiv preprint arXiv:2302.08639* (2023).

[27] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5329–5333. https://doi.org/10.1109/ICASSP.2018.8461375

[28] Deyan Tang, Siwang Zhou, Hongbo Jiang, Haowen Chen, and Yonghe Liu. 2022. Gender-adversarial networks for face privacy preserving. *IEEE Internet of Things Journal* 9, 18 (2022), 17568–17576.

[29] Philipp Terhörst, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. 2019. Suppressing gender and age in face templates using incremental variable elimination. In *2019 International Conference on Biometrics (ICB)*. IEEE, 1–8.

[30] Philipp Terhörst, Kevin Riehl, Naser Damer, Peter Rot, Blaz Bortolato, Florian Kirchbuchner, Vitomir Struc, and Arjan Kuijper. 2020. PE-MIU: A training-free privacy-enhancing face recognition approach based on minimum information units. *IEEE Access* 8 (2020), 93635–93647.

[31] Tatjana Wingarz, Marta Gomez-Barrero, Christoph Busch, and Mathias Fischer. 2022. Privacy-Preserving Convolutional Neural Networks Using Homomorphic Encryption. In *2022 International Workshop on Biometrics and Forensics (IWBF)*. 1–6. https://doi.org/10.1109/IWBF55382.2022.9794535

[32] Syed Rohit Zaman, Dipan Sadekeen, M Aqib Alfaz, and Rifat Shahriyar. 2021. One Source to Detect them All: Gender, Age, and Emotion Detection from Voice. In *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*. 338–343. https://doi.org/10.1109/COMPSAC51774.2021.00055

[33] Keyu Zhu, Pascal Van Hentenryck, and Ferdinando Fioretto. 2021. Bias and variance of post-processing in differential privacy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11177–11184.

[34] Alexander Ziller, Dmitrii Usynin, Rickmer Braren, Marcus Makowski, Daniel Rueckert, and Georgios Kaissis. 2021. Medical imaging deep learning with differential privacy. *Scientific Reports* 11, 1 (2021), 1–8.