applied sciences

MDPI

Article

GANBA: Generative Adversarial Network for Biometric Anti-Spoofing

Alejandro Gomez-Alanis *, Jose A. Gonzalez-Lopez and Antonio M. Peinado

Department of Signal Theory, Telematics and Communications, University of Granada, 18010 Granada, Spain; joseangl@ugr.es (J.A.G.-L.); amp@ugr.es (A.M.P.)

* Correspondence: agomezalanis@ugr.es

Abstract: Automatic speaker verification (ASV) is a voice biometric technology whose security might be compromised by *spoofing* attacks. To increase the robustness against *spoofing* attacks, presentation attack detection (PAD) or anti-spoofing systems for detecting replay, text-to-speech and voice conversion-based *spoofing* attacks are being developed. However, it was recently shown that adversarial *spoofing* attacks may seriously fool anti-spoofing systems. Moreover, the robustness of the whole biometric system (ASV + PAD) against this new type of attack is completely unexplored. In this work, a new generative adversarial network for biometric anti-spoofing (GANBA) is proposed. GANBA has a twofold basis: (1) it jointly employs the anti-spoofing and ASV losses to yield very damaging adversarial *spoofing* attacks, and (2) it trains the PAD as a discriminator in order to make them more robust against these types of adversarial attacks. The proposed system is able to generate adversarial *spoofing* attacks which can fool the complete voice biometric system. Then, the resulting PAD discriminators of the proposed GANBA can be used as a defense technique for detecting both original and adversarial *spoofing* attacks. The physical access (PA) and logical access (LA) scenarios of the ASVspoof 2019 database were employed to carry out the experiments. The experimental results show that the GANBA attacks are quite effective, outperforming other adversarial techniques when applied in white-box and black-box attack setups. In addition, the resulting PAD discriminators are more robust against both original and adversarial *spoofing* attacks.

Keywords: adversarial attacks; automatic speaker verification (ASV); anti-spoofing; presentation attack detection (PAD); voice biometrics

Citation: Gomez-Alanis, A.; Gonzalez-Lopez , J.A.; Peinado, A.M. GANBA: Generative Adversarial Network for Biometric Anti-Spoofing. *Appl. Sci.* **2022**, *12*, 1454. https://doi.org/10.3390/app12031454

Academic Editor: Francesc Alías

Received: 5 January 2022
Accepted: 28 January 2022
Published: 29 January 2022

1. Introduction

Biometric authentication [1] aims to authenticate the identity claimed by a given individual based on the samples measured from biological characteristics (e.g., voice, face, and fingerprints). In recent years, however, automatic speaker verification (ASV) technology has shown vulnerability to security attacks where impostors try to fraudulently access the system by inputting speech similar to the voice of a genuine user [2,3]. These security threats for voice biometric systems are known as *spoofing* attacks.

Four types of *spoofing* attacks were identified by the scientific community [4]: (i) replay (i.e., using a pre-recorded voice of the target user), (ii) impersonation (i.e., mimicking the voice of the target voice), or either using (iii) text-to-speech synthesis (TTS) or (iv) voice conversion (VC) systems to generate artificial speech resembling the voice of a genuine user. Moreover, these attacks can be presented to the ASV system using either logical access (LA) or physical access (PA) scenarios. In the LA scenario, the sensor is by-passed and TTS- or VC-based attacks are directly injected into the ASV system. In the PA attack scenario, the replayed *spoofing* signal is presented to or captured by the sensor, i.e., the microphone.

Anti-spoofing or presentation attack detection (PAD in ISO/IEC 30107 nomenclature [5]) for voice biometrics has gained increased attention in recent years as shown by the organization of multiple evaluation challenges: (i) ASVspoof 2015 [6], which focused on

detecting TTS- and VC-based *spoofing* attacks; (ii) BTAS 2016 [7], which addressed both the detection of PA and LA-based attacks; (iii) ASVspoof 2017 [8], which focused on detecting real replay *spoofing* attacks under noisy environments; (iv) ASVspoof 2019 [9], which addressed both the detection of simulated replay attacks and LA-based attacks generated with the latest TTS and VC technologies; and (v) ASVspoof 2021 [10] which addressed the same LA-based attacks as the ASVspoof 2019 Challenge but communicated across telephony and VoIP networks with various coding and transmission effects. It also addressed PA-based attacks in real and physical spaces.

The need to strengthen voice biometric systems [11] against *spoofing* attacks, has boosted the development of anti-spoofing or PAD systems capable of detecting *spoofing* speech [12–14]. In the last ASVspoof challenges [8–10], the complete voice biometric system which was evaluated is a cascaded integration of ASV and PAD systems based on score-level decisions, as depicted in Figure 1.

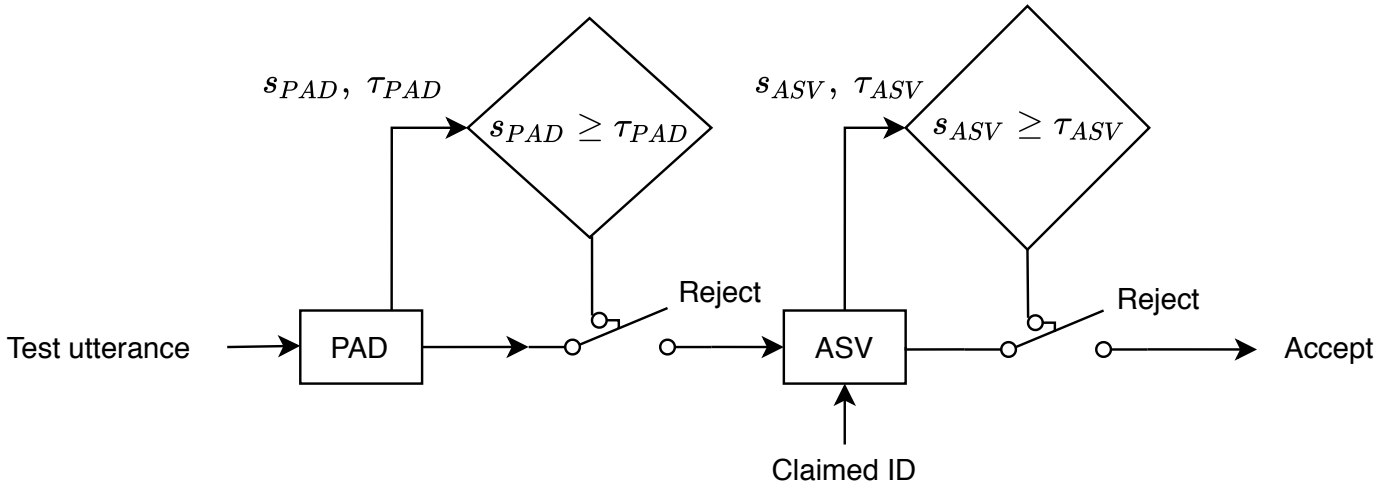

Figure 1. Block diagram of a cascade PAD + ASV voice biometric system based on score-level decisions. τ_{ASV}, s_{ASV} and τ_{PAD}, s_{PAD} are the ASV/PAD thresholds and scores.

More recently, different investigations [15–17] showed that anti-spoofing systems are also vulnerable to adversarial attacks [18]. This type of attack perturbs benign samples, normally in a way imperceptible to humans, which are able to fool deep neural network (DNN)-based models [19]. There are two main categories of adversarial attacks: (i) white-box adversarial attacks, and (ii) black-box adversarial attacks. In this paper, those adversarial attacks where the attacker has all the information about the victim model (i.e., its artifacts, weights, and model architecture) will be referred to as white-box attacks. Similarly, the term black-box is used to indicate those attacks where the attacker does not have knowledge about the victim model but the attacker can repeatedly query it to obtain a student model (surrogate) of the teacher (victim). In order to do this, the attacker uses the binary outputs provided by the victim model (acceptance/rejection) as ground-truth labels.

The main contributions proposed in this work are summarized in the following:

- We study the robustness of the complete voice biometric system against adversarial *spoofing* attacks.
- We also propose a novel generative adversarial network for biometric anti-spoofing (GANBA) which generates adversarial *spoofing* attacks capable of fooling the PAD system without being detected by the ASV system, i.e., without changing the speaker information of the utterance. Moreover, while our previous work [15] was focused on adversarial attack generation, here we also train the PAD discriminator so that it provides us with a reinforced defense against adversarial and even original *spoofing* attacks.
- To the best of our knowledge, adversarial *spoofing* attacks have only been studied on TTS and VC *spoofing* attacks (LA scenarios). In this paper, replay *spoofing* attacks (PA scenarios) are also considered.

The rest of this paper is organized as follows. Section 2 outlines the PAD, ASV, and complete voice integration systems, as well as the metrics to evaluate them. In addition, adversarial *spoofing* attacks as well as some well-known adversarial attacks, employed as baselines in this work, are discussed. The GANBA framework for both white-box and black-box scenarios are proposed in Section 3. Then, Section 4 describes the systems details, speech corpora, and the metrics evaluated in the experiments. Section 5 discusses the experimental results. Finally, Section 6 summarizes the conclusions derived from this research.

2. Background

This section is devoted to briefly describe the existing standalone PAD and ASV approaches, as well as the complete voice integration systems including the metrics to evaluate all of them. Moreover, Section 2.4 provides a detailed description of adversarial *spoofing* attacks.

2.1. Automatic Speaker Verification (ASV)

An ASV system is able to determine whether an utterance is uttered by the claimed speaker or not. In order to do it, it typically obtains the speaker information of the utterance by extracting either i-vector [20] or x-vector features [21]. In the verification phase, the ASV system extracts the feature vectors of the enrollment and test utterances, and they are usually mapped into a more discriminative subspace using, for example, linear discriminant analysis (LDA). Then, the ASV score of the test utterance is typically obtained using one of the following techniques:

- Probabilistic Linear Discriminant Analysis (PLDA) [22,23]: it is a probabilistic framework which is able to model the inter- and intra-speaker variability. There are three types of PLDA models [24]: simplified [25], standard [22], and two-covariance [26]. In all variants, the expectation-maximization (EM) algorithm [27] is used to train the PLDA model.
- B-vector [28]: it is a DNN-based model which considers ASV as a binary classification problem. Specifically, from the x-vectors $\boldsymbol{x}_{\text{enroll}}$ and $\boldsymbol{x}_{\text{test}}$ computed for each pair of utterances (enrollment and test utterances), a b-vector representing the relationship between $\boldsymbol{x}_{\text{enroll}}$ and $\boldsymbol{x}_{\text{test}}$ is computed as follows,

$$\boldsymbol{b} = \left[\boldsymbol{x}_{\text{enroll}} \oplus \boldsymbol{x}_{\text{test}}, \left|\boldsymbol{x}_{\text{enroll}} \ominus \boldsymbol{x}_{\text{test}}\right|, \boldsymbol{x}_{\text{enroll}} \otimes \boldsymbol{x}_{\text{test}}\right], \tag{1}$$

  where $\oplus$, $\ominus$ and $\otimes$ are the element-wise addition, subtraction and multiplication operations, respectively. Then, the b-vector features are fed to a binary DNN which determines whether the enrollment and test utterances are uttered by the same or different speaker.

An ASV system is typically evaluated on a test dataset which contains utterances uttered by either *bonafide* target speakers or *zero-effort* impostors [29]. The equal error rate (EER) is the most common metric to evaluate it, which is the operating point at which the false rejection rate (FRR) equals the false acceptance rate (FAR). However, the EER metric does not account for either the costs of falsely accepting impostors or missing target users, nor the prior probabilities of each. In order to take these costs and priors into account, the detection cost function (DCF) metric [30] has been proposed and evaluated in the most popular speaker recognition challenges [31].

2.2. Anti-Spoofing or Presentation Attack Detection (PAD)

The goal of anti-spoofing is to differentiate between *bonafide* and *spoofing* speech. Two hypotheses are computed for each test utterance: (i) it is *bonafide* speech, or (ii) it is a spoofing attack.

In the last ASVspoof challenges [8–10], DNN-based models have been the most effective approach to differentiate between *bonafide* and *spoofing* speech. A wide range of

features have been proposed for training these models, such as linear frequency cepstral coefficients (LFCC) [32], spectrograms [33], constant Q cepstral coefficients (CQCC) [34], and raw speech samples [35].

Anti-spoofing systems are typically evaluated using the ERR_{PAD} metric, where false acceptance happens when a *spoofing* utterance is detected as a *bonafide* utterance while false rejection occurs when a *bonafide* utterance is detected as a *spoofing* attack. Moreover, in order to take costs and priors of the different hypotheses into account, the ASV-constrained minimum tandem detection cost function (min-tDCF) metric [36] has been recently proposed to evaluate anti-spoofing systems. This has been the primary metric of the ASVspoof 2019 and 2021 challenges [9,10].

2.3. Voice Integration Systems: Joint ASV and PAD

In the integration approach, each utterance has two attributes: (i) an indicator of the target speaker ($\mathcal{S}$), and (ii) an indicator of the *bonafide* speech ($\mathcal{N}$). Therefore, the null hypothesis $\mathcal{H}_{(\mathcal{S},\mathcal{N})}$ is that the test utterance is *bonafide* speech uttered by the target speaker. On the other hand, the complementary hypotheses is a union of the other three hypotheses:

$$\mathcal{H}_{\overline{(\mathcal{S},\mathcal{N})}} = \mathcal{H}_{(\mathcal{S},\overline{\mathcal{N}})} \cup \mathcal{H}_{(\overline{\mathcal{S}},\mathcal{N})} \cup \mathcal{H}_{(\overline{\mathcal{S}},\overline{\mathcal{N}})}, \tag{2}$$

where $(\mathcal{S},\overline{\mathcal{N}})$ denotes a *spoofing* attack, $(\overline{\mathcal{S}},\mathcal{N})$ represents *bonafide* speech uttered from a non-target speaker (i.e., it is *zero-effort* impostor), and $(\overline{\mathcal{S}},\overline{\mathcal{N}})$ represents *spoofing* speech from a non-target speaker. The latter case, commonly referred to as naive attack, does not make much sense in an authentication context and it is usually discarded. Then, there are three types of utterances that PAD and ASV systems may encounter: (i) *bonafide* or *genuine target*, (ii) *zero-effort impostor* or *genuine non-target*, and (iii) *spoofing target* attacks.

The integration of PAD and ASV systems can be achieved at the score level [37] or at the feature level [11,38]. Most existing integration methods perform the integration at the score level, where dedicated classifiers are developed for both PAD and ASV separately, and the scores computed by each separate system are combined. In this work, we focus at the score-level integration. Specifically, we use the cascaded integration system depicted in Figure 1 which has been used in the last three ASVspoof challenges [8–10].

The integration systems are typically evaluated using the $\text{EER}_{\text{joint}}$ which can be measured, for example, on a test dataset that contains a combination of *bonafide* utterances, *zero-effort* attacks and *spoofing* attacks. However, the $\text{EER}_{\text{joint}}$ does not account for the costs of falsely accepting *spoofing* attacks and *zero-effort* impostors or missing target genuine users, nor the prior probabilities of each. In order to take these costs and priors into account, the min-tDCF metric [36,39] was recently proposed for evaluating complete voice integration systems based on score-level decisions.

2.4. Adversarial Spoofing Attacks

Adversarial *spoofing* attacks can be generated by adding a minimally perceptible perturbation to the input *spoofing* utterance. The core idea of this type of adversarial attack is to refine the original *spoofing* attack so that it is more difficult to be detected by the PAD system. In other words, the goal of adversarial *spoofing* attacks is to fool the anti-spoofing or PAD system by maximizing the *bonafide* class likelihood with respect to that of the *spoofing* class.

Adversarial *spoofing* attacks can be generated by freezing the parameters $\boldsymbol{\theta}$ of the DNN-based anti-spoofing model and performing a gradient descent algorithm which is able to update the input spectrum features $\boldsymbol{X}$ of the *spoofing* utterance so that the PDA misclassifies it as *bonafide*. Mathematically, it is an optimization problem which tries to find a sufficiently small perturbation δ which satisfies:

$$\begin{aligned} \tilde{\boldsymbol{X}} &= \boldsymbol{X} + \delta, \\ f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}) &= \tilde{y}, \\ f_{\boldsymbol{\theta}}(\boldsymbol{X}) &= y, \end{aligned} \tag{3}$$

where $\tilde{\boldsymbol{X}}$ denotes the perturbed input spectrum features, f denotes a well-trained DNN-based anti-spoofing model parameterized by $\boldsymbol{\theta}$, δ is the additive spectrum perturbation, y is the ground-truth label corresponding to the original input spectrum features $\boldsymbol{X}$, and $\tilde{y}$ denotes the label of the targeted class of the adversarial *spoofing* attack, i.e., the label of the *bonafide* class. Normally, the subspace of allowed perturbations is denoted by $\boldsymbol{\Delta}$, so that the perturbation $\delta \in \boldsymbol{\Delta}$. In other words, the manipulative capability of the adversarial attack is formalized by the subspace $\boldsymbol{\Delta}$. Typically, $\boldsymbol{\Delta}$ is a small l_∞-norm sphere ($\boldsymbol{\Delta} = \{\delta \mid \|\delta\|_\infty \leq \epsilon\}$, $\epsilon \geq 0 \in \mathbb{R}$).

Two of the most popular adversarial attack techniques are: (i) gradient sign method (FGSM) [40], and (ii) projected gradient descent (PGD) [41]. The FGSM attack generates the perturbation δ by taking a single step toward the direction pointed by the gradient as follows,

$$\delta = \epsilon \cdot \text{sign}(\nabla_{\boldsymbol{X}} \text{Loss}(\boldsymbol{\theta}, \boldsymbol{X}, y)), \tag{4}$$

where sign is an operation which takes the sign of its argument, and Loss is the loss function of the well-trained PAD neural network whose parameters are denoted by $\boldsymbol{\theta}$. Unlike FGSM, implemented as a single-step procedure, PGD is iterative. Thus, initializing with the original input *spoofing* spectrum features $\boldsymbol{X}_0 = \boldsymbol{X}$, the spectra of the *spoofing* attack is iteratively updated as follows,

$$\boldsymbol{X}_{n+1} = \text{clip}(\boldsymbol{X}_n + \alpha \cdot \text{sign}(\nabla_{\boldsymbol{X}} Loss(\boldsymbol{\theta}, \boldsymbol{X}, y)),\ \forall\, n = 0, ..., N-1, \tag{5}$$

where $n = 0, ..., N-1$ denotes the iteration index (up to N iterations), and clip denotes a function which applies element-wise clipping such that $\|\boldsymbol{X}_n - \boldsymbol{X}\|_\infty \leq \epsilon$, $\epsilon \geq 0 \in \mathbb{R}$.

3. Proposed Method

In this work, we propose a generative adversarial network for biometric anti-spoofing (GANBA) in order to generate adversarial *spoofing* attacks and, at the same time, train the PAD discriminator in order to make it more robust against this type of attack. The generator of the proposed GANBA is a neural network which is in charge of transforming the original input *spoofing* spectrum features into adversarial *spoofing* spectrum features against a target voice biometric system. Thus, the discriminator of the GANBA is a complete voice biometric system (ASV + PAD) which tries to differentiate between *bonafide* and *spoofing* speech (PAD system), and verify the identity of the enrolled speakers (ASV system).

The PAD and ASV models of the proposed GANBA provides either a probability distribution across the *bonafide* and *spoofing* class labels (white-box scenario) or just a binary decision indicating whether the test utterance is accepted or rejected by the biometric system (black-box scenario). In both scenarios, the goal of the proposed GANBA generator is to provide high quality adversarial *spoofing* attacks from *spoofing* speech able to fool the anti-spoofing system while undetected by the ASV subsystem (that is, the speaker information contained in the utterance is not modified). In contrast, the objective of the GANBA discriminator (complete voice biometric system) is to detect both the original (We refer to the original *spoofing* attacks as those unseen *spoofing* attacks of the test dataset which are not modified by any adversarial perturbation) and adversarial *spoofing* attacks.

3.1. White-Box GANBA

Figure 2 depicts the proposed GANBA architecture for the white-box scenario. The inputs to the GANBA generator are the short-time Fourier transform (STFT) features of a *spoofing* utterance, so that it modifies its spectra in order to refine the *spoofing* attack. The output of the GANBA generator is fed to the PAD and ASV subsystems of the target biometric system. The ASV system only consists of a time-delay neural network (TDNN) [21] for x-vector extraction (the only component of the ASV system needed in the white-box scenario). This feature extractor is fed with the Mel-frequency cepstral coefficients (MFCCs) of the corresponding utterance obtained through the log-power magnitude spectrum features (STFT) extracted previously, as shown in the diagram of Figure 2. On the other hand, the PAD

system based on DNNs is also fed with the STFT features of the corresponding utterance and it provides the softmax output vector of the utterance, whose first component indicates the probability of the utterance being *bonafide*. The goal of the proposed framework is to train a GANBA generator capable of generating adversarial *spoofing* attacks which can fool the anti-spoofing system while not causing any changes to the ASV x-vector output, i.e., the adversarial attacks should not change the feature x-vector since it contains the speaker information of the utterance.

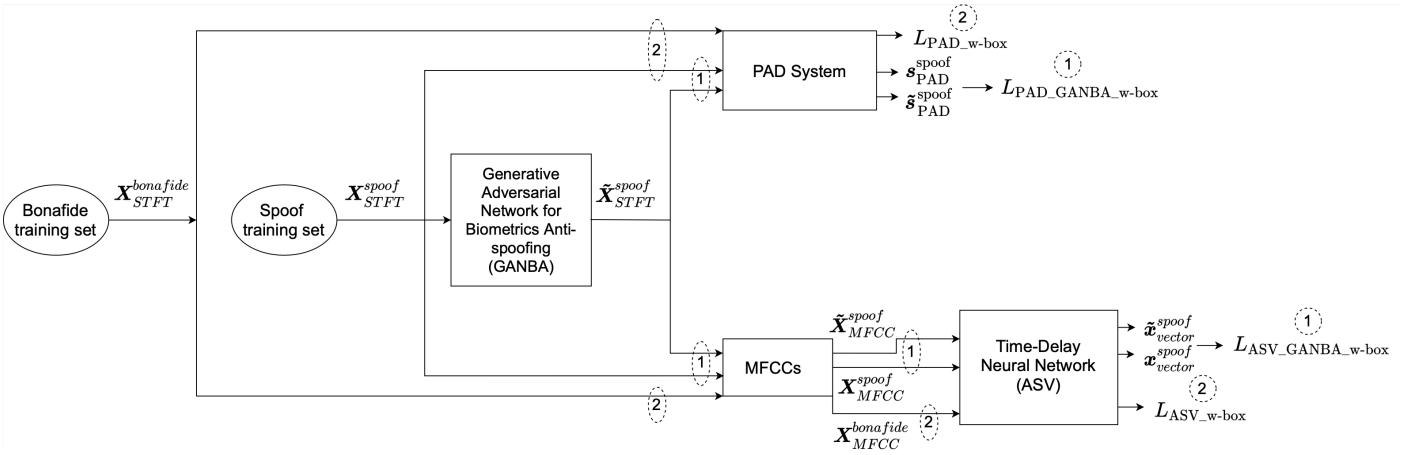

Figure 2. Generative adversarial network for biometric anti-spoofing (GANBA) framework for white-box scenarios. Step 1: generator-only training (ASV and PAD parameters frozen, with encircled outputs corresponding to Equations (7) and (8), respectively). Step 2: discriminator (ASV + PAD) training (encircled outputs corresponding to classical cross-entropy loss function).

As is shown in Figure 2, there are two different steps for training the whole architecture. Step 1 denotes the training of the GANBA generator, where the PAD and ASV parameters are not modified but gradients are computed and back-propagated to the GANBA generator. At this stage, the *spoofing* speech dataset is used only to train the GANBA generator. On the other hand, Step 2 denotes the training of the biometric system, PAD and ASV components, which makes up the discriminator of the GANBA framework. While the TDNN of the ASV system is trained using only the *bonafide* speech dataset, the PAD system is trained using both the *bonafide* and *spoofing* speech datasets.

Step 1 is in charge of optimizing the GANBA generator parameters in the white-box (w-box) scenario. In particular, the following loss function is minimized:

$$\begin{aligned} L_{\text{GANBA\_w-box}} &= L_{\text{PAD\_GANBA\_w-box}}\left(\boldsymbol{s}_{\text{PAD}}^{\text{spoof}}, \tilde{\boldsymbol{s}}_{\text{PAD}}^{\text{spoof}}\right) \\ &+ \beta \cdot L_{\text{ASV\_GANBA\_w-box}}\left(\boldsymbol{x}_{\text{vector}}^{\text{spoof}}, \tilde{\boldsymbol{x}}_{\text{vector}}^{\text{spoof}}\right), \end{aligned} \tag{6}$$

where

$$L_{\text{ASV\_GANBA\_w-box}}\left(\boldsymbol{x}_{\text{vector}}^{\text{spoof}}, \tilde{\boldsymbol{x}}_{\text{vector}}^{\text{spoof}}\right) = \left\| \boldsymbol{x}_{\text{vector}}^{\text{spoof}} - \tilde{\boldsymbol{x}}_{\text{vector}}^{\text{spoof}} \right\|_2, \tag{7}$$

$$L_{\text{PAD\_GANBA\_w-box}}\left(\boldsymbol{s}_{\text{PAD}}^{\text{spoof}}, \tilde{\boldsymbol{s}}_{\text{PAD}}^{\text{spoof}}\right) = \left\| r_\alpha\left(\boldsymbol{s}_{\text{PAD}}^{\text{spoof}}\right) - \tilde{\boldsymbol{s}}_{\text{PAD}}^{\text{spoof}} \right\|_2. \tag{8}$$

$L_{\text{ASV\_GANBA\_w-box}}$ and $L_{\text{PAD\_GANBA\_w-box}}$ are the loss functions associated with the ASV and PAD systems, respectively, and β is a hyper-parameter which weights the relative importance of these two losses. $\boldsymbol{x}_{\text{vector}}^{\text{spoof}}$ and $\tilde{\boldsymbol{x}}_{\text{vector}}^{\text{spoof}}$ denote the x-vectors of the original and adversarial *spoofing* utterances, respectively. Likewise, vectors $\boldsymbol{s}_{\text{PAD}}^{\text{spoof}}$ and $\tilde{\boldsymbol{s}}_{\text{PAD}}^{\text{spoof}}$ represent the output probability sets provided by the PAD system for the original and adversarial

spoofing utterances, respectively. Moreover, r_α is a re-ranking function which is formulated as follows,

$$r_\alpha\left(\boldsymbol{s}_{\text{PAD}}^{\text{spoof}}\right) = \text{norm}\left(\begin{cases} \alpha \cdot \max\left(\boldsymbol{s}_{\text{PAD}}^{\text{spoof}}\right) & k = 0 \\ \boldsymbol{s}_{\text{PAD}}^{\text{spoof}}(k) & k \neq 0 \end{cases}\right), \tag{9}$$

where k is the index class variable of the $\boldsymbol{s}_{\text{PAD}}^{\text{spoof}}$ probability vector, with $k = 0$ representing the class of *bonafide* speech, $\alpha > 1$ is an additional hyper-parameter which defines how large $\boldsymbol{s}_{\text{PAD}}^{\text{spoof}}(k = 0)$ (i.e., the *bonafide* class probability) is with respect to the current maximum probability class, and norm denotes a normalizing function which provides a valid probability distribution.

On the other hand, Step 2 indicates the training of the voice biometric system (ASV + PAD) which acts as the GAN discriminator. The ASV system based on a TDNN is trained as a classifier using only the *bonafide* speech dataset. Likewise, the PAD system is also trained as a classifier using both the *spoofing* and *bonafide* speech datasets. Finally, Step 1 and Step 2 follow each other to train the proposed GANBA as a generative adversarial network (GAN), where the GANBA generator is in charge of producing the adversarial *spoofing* attacks and is trained during Step 1, while the PAD and ASV discriminators are in charge of detecting the voice biometric attacks and are trained during Step 2 as normal ASV and PAD systems using their corresponding loss functions [2,21].

3.2. Black-Box GANBA

Figure 3 depicts the proposed GANBA architecture for the black-box scenario. Similar to the white-box case, the goal of the proposed system is the generation of adversarial *spoofing* attacks capable of fooling the target PAD system (teacher PAD) and bypassing the target ASV system (teacher ASV) by not modifying the x-vector representation which encodes the speaker information of the utterance. However, the main limitation of the black-box scenario is that the attacker does not have access to the target system (teacher) parameters.

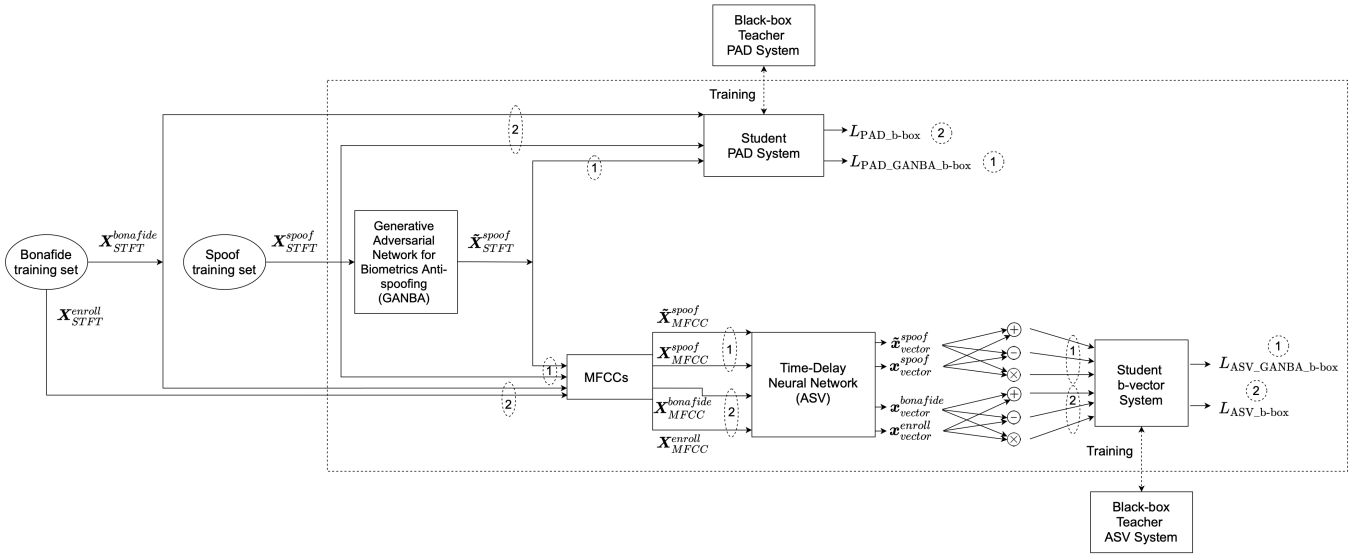

Figure 3. Generative adversarial network for biometric anti-spoofing (GANBA) framework for black-box scenarios. Step 1: generator-only training (ASV and PAD parameters frozen, with encircled outputs corresponding to Equations (11) and (12), respectively). Step 2: discriminator (ASV + PAD) training (encircled outputs corresponding to classical cross-entropy loss function).

In the proposed GANBA framework, the attacker makes requests to the black-box target (teacher) biometric system but only obtains a binary decision response of acceptance/rejection. The binary response is considered to be the ground-truth label for training

the student PAD and b-vector [28] systems of the GANBA discriminator. It is worth noticing that we assume that a rejection of the teacher system is made by the teacher PAD system since we make sure that the x-vector representation of the utterance is minimally perturbed by the adversarial *spoofing* attack. Thus, the student b-vector and PAD systems are trained as binary classifiers able to imitate the behavior of the teacher ASV and PAD systems, respectively. In particular, the student b-vector system computes the probability that the two input x-vectors represent the same speaker, i.e., that $P(b(\boldsymbol{x}_{\text{vector}}, \tilde{\boldsymbol{x}}_{\text{vector}}) = 1)$, with b denoting the b-vector model.

As shown in Figure 3, there are again two different steps for training the whole architecture. Step 1 denotes again the training of the GANBA generator, where the ASV and PAD network parameters are not modified but the gradients are back-propagated to the GANBA generator. Thus, Step 1 is in charge of optimizing the parameters of the GANBA generator in the black-box (b-box) case by minimizing the following loss function:

$$\begin{aligned} L_{\text{GANBA\_b-box}} &= L_{\text{PAD\_GANBA\_b-box}}(\tilde{\boldsymbol{s}}_{\text{PAD}}^{\text{spoof}}) \\ &+ \beta \cdot L_{\text{ASV\_GANBA\_b-box}}(\boldsymbol{x}_{\text{vector}}^{\text{spoof}}, \tilde{\boldsymbol{x}}_{\text{vector}}^{\text{spoof}}), \end{aligned} \tag{10}$$

where

$$L_{\text{ASV\_GANBA\_b-box}}(\boldsymbol{x}_{\text{vector}}^{\text{spoof}}, \tilde{\boldsymbol{x}}_{\text{vector}}^{\text{spoof}}) = 1 - P(b(\boldsymbol{x}_{\text{vector}}^{\text{spoof}}, \tilde{\boldsymbol{x}}_{\text{vector}}^{\text{spoof}}) = 1), \tag{11}$$

$$L_{\text{PAD\_GANBA\_b-box}}(\tilde{\boldsymbol{s}}_{\text{PAD}}^{\text{spoof}}) = \left\| \text{onehot}(k = 0) - \tilde{\boldsymbol{s}}_{\text{PAD}}^{\text{spoof}} \right\|_2. \tag{12}$$

$L_{\text{ASV\_GANBA\_b-box}}$ and $L_{\text{PAD\_GANBA\_b-box}}$ are the loss components associated with the ASV and PAD systems, respectively. Furthermore, onehot is the one-hot function [42] and $k = 0$ is the *bonafide* class index. Using this function, the input *spoofing* utterance is presented as *bonafide* and the PAD is fooled.

On the other hand, Step 2 denotes the training of the student biometric system (ASV b-vector + PAD) which acts as the discriminator of the GAN. The TDNN employed for x-vector extraction is pretrained and its parameters are also frozen in this step. However, the b-vector system is trained as a binary classifier [28] employing the test and enrollment utterances from the *bonafide* speech dataset. Likewise, the PAD system is also trained as a binary classifier using both *bonafide* and *spoofing* utterances. In both cases, the ground truth labels are taken from the binary responses of the black-box target/teacher biometric system. Similar to the white-box scenario, Step 1 and Step 2 follow each other in order to train the proposed GANBA as a typical GAN.

4. Experimental Setup

In this section, we describe the databases, spectral analysis, implementation details, and evaluation metrics employed in the experiments.

4.1. Speech Datasets

We used the ASVspoof 2019 corpus [43] to train and evaluate all the systems. This database is split into two subsets to allow PA and LA evaluation. Moreover, it does not only include protocols for evaluating PAD systems, but also for evaluating ASV and ASV+PAD integration systems. First, we employed this database for training the standalone anti-spoofing systems in the PA and LA scenarios, respectively. Then, we only used the *spoofing* utterances for generating adversarial *spoofing* attacks in order to bypass the complete voice biometric system. It is worth noticing that the adversarial examples were not generated from *bonafide* utterances because we consider that doing so they would lose their *bonafide* character.

To train the TDNN [21] of the ASV system as an x-vector features extractor, we also employed the Voxceleb1 database [44] which contains more than 1000 speakers. Moreover, in order to train the b-vector [28] ASV scoring system in the black-box scenario, the *bonafide* utterances (from the ASVspoof 2019 and Voxceleb1 development datasets) were used, thus following the training protocol described in [11].

4.2. Spectral Analysis

The PAD systems were fed with log-power short-time Fourier transform (STFT) features with 256 frequency bins and 600 acoustic frames. In order to obtain the STFT features, a Hanning analysis window with a 10 ms frame shift and 25 ms of window length was employed. On top of these STFT features, 24 MFCCs (including the $C0$ cepstral coefficient), obtained with the Kaldi recipe [45], were extracted to feed the TDNN-based ASV system.

4.3. Implementation Details

Three state-of-the-art anti-spoofing systems were adapted from other works: a light convolutional neural network (LCNN) [2], a residual neural network (ResNet34) [46] and a Squeeze-Excitation network (SENet50) [46]. The softmax layer output of the DNN-based models was directly used to obtain the PAD scores. For ASV, a TDNN model for x-vector feature extraction [21] was trained. Then, two ASV scoring systems were trained: (i) a standard PLDA [22]; and (ii) a b-vector system [28].

The generator of the proposed GANBA framework is a convolutional neural network (CNN) with five convolutional layers (16, 32, 48, 48, and 3 channels). Furthermore, it uses a kernel size of 3×3 as well as leaky ReLU activations. The Adam optimizer [47] (with learning rate 3×10^{-4}) was used to train the GANBA generator. Moreover, a grid search across the development dataset of the ASVspoof 2019 database was used in order to find the best empirical values of the hyper-parameters $\alpha = 10$ and $\beta = 0.001$. All the deep learning frameworks were trained using the Pytorch toolkit [48].

The PGD method uses $N = 30$ training iterations for generating the adversarial *spoofing* attack, and for evaluation the number of iterations is set to $N = 100$ [49]. The magnitude of the perturbation is configured with the ϵ parameter. In the experiments, we do a grid search between $\epsilon = 0.1$ and $\epsilon = 5.0$ in order to find the optimal perturbation of the FSGM and PGD techniques. However, the magnitude of the GANBA perturbation is not restricted to any specific value. This is one of the main advantages of the proposed GANBA technique since it is in charge of finding the optimal perturbation value by itself in order to fool the PAD system without being detected by the ASV system, i.e., without changing the speaker information of the utterance.

4.4. Evaluation Metrics

A specific EER (EER_{ASV}) was used for ASV. We evaluated this metric either including only *bonafide* utterances or including both *bonafide* and *spoofing* utterances. Likewise, the PAD systems were also evaluated using the EER_{PAD} across all *spoofing* attacks. To compute the performance of the complete voice integration system, any utterance rejected by either the ASV or PAD systems was arbitrarily assigned a $-\infty$ score. Finally, the integration systems were evaluated using the min-tDCF [39] metric and the joint EER ($\text{EER}_{\text{joint}}$) with the same configuration as that of the ASVspoof 2019 challenge [9]. The ASVspoof 2019 test datasets were used to evaluate all the ASV, PAD and complete voice integration systems.

5. Results

This section presents the experimental results from the evaluation of the described techniques on the ASVspoof 2019 corpus. First, Section 5.1 presents the results of different biometric systems without being exposed to any adversarial *spoofing* attacks. Then, Section 5.2 evaluates the vulnerability of a biometric system to white-box adversarial *spoofing* attacks. Likewise, Section 5.2 is devoted to the evaluation of the black-box adversarial *spoofing* attacks, where the details of the target biometric system remain unknown to the attacker. In both Sections 5.2 and 5.3, the proposed white-box and black-box GANBA attacks will be compared to other classical adversarial *spoofing* attacks, respectively. Finally, Section 5.4 presents the results of the biometric system after applying two defense techniques: (i) adversarial training of the PAD discriminator using the generated adversarial *spoofing* attacks, and (ii) using the PAD discriminator trained within the proposed GANBA framework.

5.1. Voice Biometric Systems Results

Table 1 presents the baseline results of six biometric systems which consist of the combination of three PAD (LCNN, SENet50 and ResNet34) and two ASV (TDNN + PLDA and TDNN + b-vector) systems. These biometric systems are not still exposed to any adversarial *spoofing* attacks. These PA and LA anti-spoofing systems have been shown to provide some of the best single PAD performance in the ASVspoof 2019 challenge [9]. The best biometric system is the combination of LCNN and TDNN + PLDA as the PAD and ASV systems, respectively. Although the ASV system provides EER of 6.87% and 4.71% over the PA and LA datasets, respectively, when evaluating exclusively with *bonafide* utterances (target and non-target), its performance meaningfully degrades when *spoofing* utterances are also evaluated (in particular, 18.43% and 30.58% in the PA and LA evaluation datasets, respectively). This TDNN + PLDA / LCNN biometric system will be used as the teacher system for the black-box scenario in Section 5.3.

5.2. White-Box Attacks Results

Figure 4 shows the $\text{EER}_{\text{joint}}$ of the best TDNN + PLDA/LCNN based biometric system evaluated in the previous section when being exposed to white-box adversarial attacks. The PGD technique as expected achieves slightly better results than the FGSM technique since PGD uses an iterative procedure for generating adversarial *spoofing* attacks. However, the proposed GANBA attacks outperform the other adversarial attacks, obtaining 20.94% and 27.63% higher absolute $\text{EER}_{\text{joint}}$ with respect to the best PGD configuration (i.e., $\epsilon = 1.0$) in the PA and LA evaluation datasets, respectively. Another remarkable result is that using a hyper-parameter ϵ higher than 2.0 in PGD and FGSM, the perturbation is effectively detected by the biometric system. In such cases, the adversarial *spoofing* attacks may perform even worse than when only using the original *spoofing* attacks, i.e., when not generating adversarial attacks from *spoofing* speech (denoted by 'No Processing').

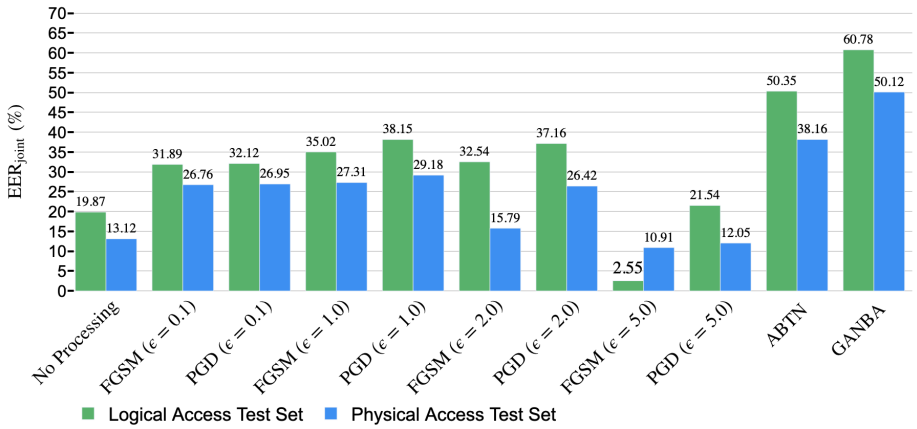

Figure 4. $\text{EER}_{\text{joint}}$(%) of the white-box adversarial *spoofing* attacks evaluated on the ASVspoof 2019 logical and physical access test datasets. The ASV and PAD systems are the state-of-the-art TDNN + PLDA and LCNN systems, respectively.

5.3. Black-box Attacks Results

Figures 5 and 6 show the $\text{EER}_{\text{joint}}$ of the state-of-the-art TDNN + PLDA / LCNN biometric system for the black-box scenario when it is attacked with three different student biometric systems in the PA and LA scenarios, respectively. As shown in Figure 2, the ASV system of the attacker is a TDNN + b-vector system. On the other hand, each student biometric system uses a different PAD system: ResNet34, SENet50, and LCNN. Moreover, three types of adversarial *spoofing* attacks are employed by each student biometric system: FGSM ($\epsilon = 1.0$), PGD ($\epsilon = 1.0$) and the proposed GANBA attack.

Table 1. Results of different complete voice biometric systems evaluated on the ASVspoof 2019 logical access (LA) and physical access (PA) test datasets in terms of EER_{PAD}(%), EER_{ASV}(%), EER_{joint}(%) and min-tDCF. (*) The EER_{ASV} metric is evaluated considering both *bonafide* and *spoofing* utterances.

Biometric Systems		**Logical Access Attacks**				**Physical Access Attacks**			
PAD	**ASV**	**EER_{PAD} (%)**	**EER_{ASV} (%)**	**EER_{joint} (%)**	**min-tDCF**	**EER_{PAD} (%)**	**EER_{ASV} (%)**	**EER_{joint} (%)**	**min-tDCF**
LCNN	TDNN + PLDA	5.85	4.71/30.58 *	19.87	0.1237	4.62	6.87/18.43 *	13.12	0.1221
LCNN	TDNN + b-vector	5.85	4.89/30.77 *	20.12	0.1256	4.62	7.13/19.21 *	13.89	0.1274
SENet50	TDNN + PLDA	6.29	4.71/30.58 *	21.15	0.1307	5.17	6.87/18.43 *	14.48	0.1328
SENet50	TDNN + b-vector	6.29	4.89/30.77 *	21.74	0.1332	5.17	7.13/19.21 *	14.81	0.1356
ResNet34	TDNN + PLDA	6.75	4.71/30.58 *	22.68	0.1412	5.62	6.87/18.43 *	15.75	0.1415
ResNet34	TDNN + b-vector	6.75	4.89/30.77 *	23.09	0.1456	5.62	7.13/19.21 *	15.97	0.1439

The proposed GANBA attack outperforms the best FGSM and PGD configurations ($\epsilon = 1.0$) by 27.67% and 27.08% in the LA scenario when using the LCNN PAD system. In the PA scenario, the proposed GANBA attack also outperforms them by 17.09% and 16.32% with the LCNN PAD system, respectively. It is worth noticing that the LCNN PAD system is always better than the ResNet34 and SENet50 since this system has the same architecture as the PAD system of the target (teacher) system. However, the SENet50 PAD system achieves only 3.44% and 3.09% of slower absolute $\text{EER}_{\text{joint}}$ when using the proposed GANBA attack compared to the LCNN PAD architecture, being able to effectively attack the teacher system achieving an $\text{EER}_{\text{joint}}$ of 50.12% and 40.12% in the LA and PA scenarios, respectively. This result shows how vulnerable a black-box complete voice biometric system can be to adversarial *spoofing* attacks.

Figure 7 shows some examples of the spectrogram of *bonafide* speech, *spoofing* speech generated with a replay attack of the physical access ASVspoof 2019 database, adversarial *spoofing* speech refined by the PGD method, and that refined by the proposed GANBA technique. As can be seen, it is difficult to visually differentiate between the original *spoofing* spectrogram and the generated adversarial spectrogram with the proposed method. However, our technique is able to correct some of the *spoofing* artifacts so that the replay attack is introduced in the spectrogram so that the PAD system misclassifies the generated adversarial *spoofing* utterance as a bonafide utterance. In contrast, the PGD method is not able to compensate for those artifacts, and the PAD system still classifies it as *spoofing* speech.

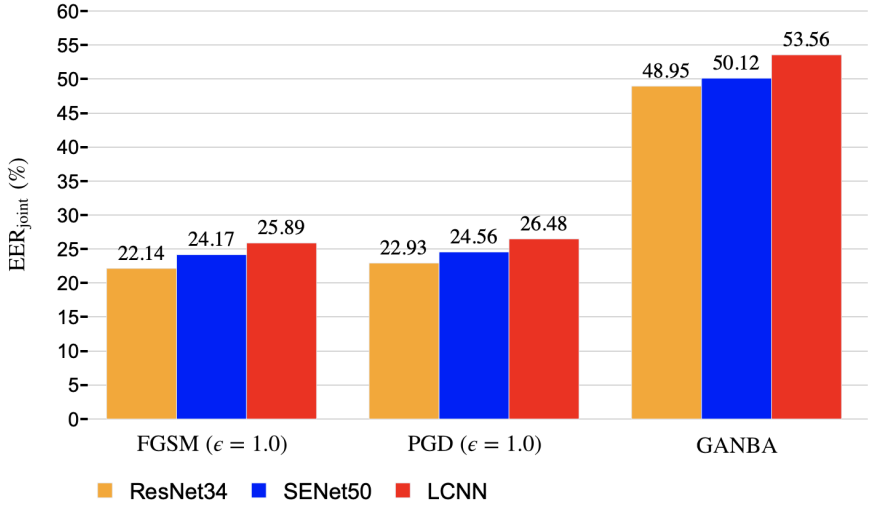

Figure 5. $\text{EER}_{\text{joint}}$(%) of the black-box adversarial *spoofing* attacks evaluated on the ASVspoof 2019 logical access (LA) test dataset with the state-of-the-art TDNN + PLDA / LCNN biometric system. There are three attackers (student systems) which use the ResNet34, SENet50 and LCNN as PAD system, respectively. The ASV system of the attacker is a TDNN + b-vector system. Every student system generates three types of attacks: FGSM ($\epsilon = 1.0$), PGD ($\epsilon = 1.0$) and the proposed GANBA attack.

5.4. Defenses against Adversarial Spoofing Attacks

Table 2 shows the performance metrics of the state-of-the-art TDNN + PLDA / LCNN biometric system when it applies two separate defense techniques: (i) adversarial training of the PAD discriminator using the generated adversarial *spoofing* attacks, and (ii) using the PAD discriminator trained within the proposed GANBA framework. Both defense techniques employ the black-box adversarial attacks generated with the SENet50 PAD system employed in Section 5.3 as discriminator. Thus, we can evaluate a realistic scenario where the target PAD system (LCNN) does not match the same architecture as that of the attacker PAD system (SENet50).

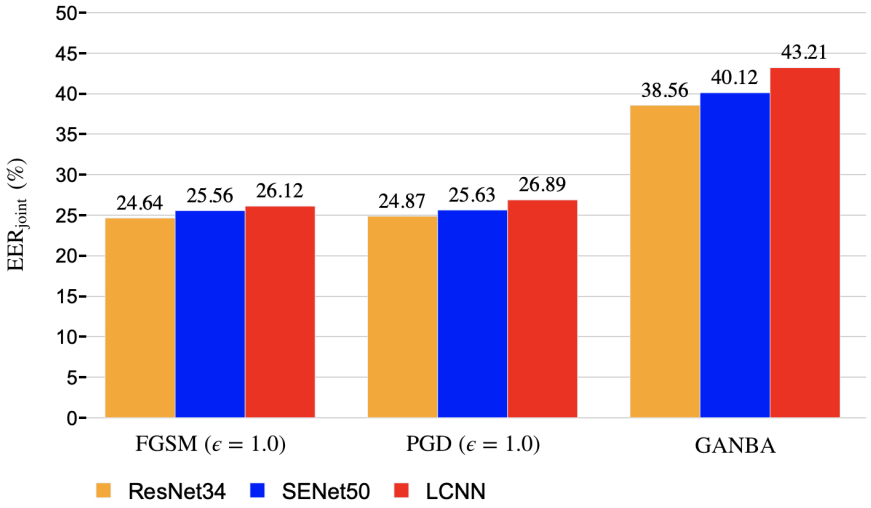

Figure 6. EER$_{joint}$(%) of the black-box adversarial *spoofing* attacks evaluated on the ASVspoof 2019 physical access (PA) test dataset with the state-of-the-art TDNN + PLDA/LCNN biometric system. There are three attackers (student systems) which use the ResNet34, SENet50 and LCNN as PAD system, respectively. The ASV system of the attacker is a TDNN + b-vector system. Every student system generates three types of attacks: FGSM ($\epsilon = 1.0$), PGD ($\epsilon = 1.0$) and the proposed GANBA attack.

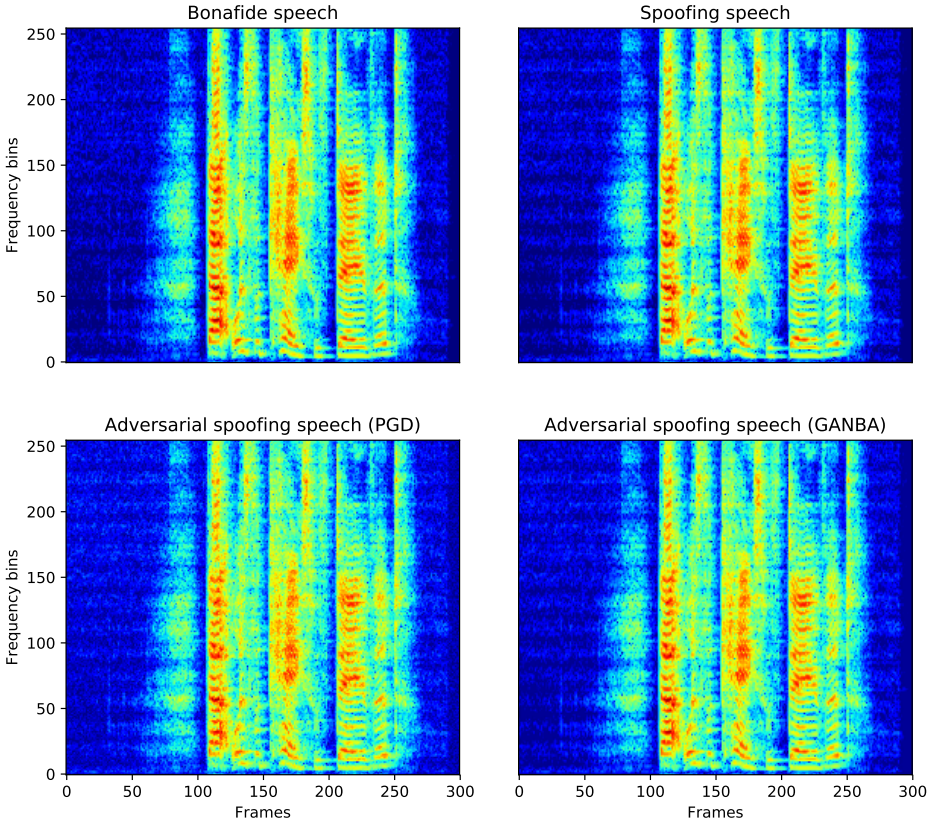

Figure 7. Spectrograms of the same utterance corresponding to: (**left upper**) bonafide speech, (**right upper**) spoofing speech generated with a replay attack, (**left bottom**) adversarial spoofing speech generated with the black-box PGD ($\epsilon = 1.0$) method, (**right bottom**) adversarial spoofing speech generated with the proposed black-box GANBA method.

Table 2 is divided into three sections separated by horizontal lines. The first row shows the performance metrics of the biometric system without being exposed to any adversarial *spoofing* attacks for the sake of comparison with the rest of attack-defense combinations. The next nine rows show the performance metrics of the biometric system evaluated after

applying a different attack-defense combination: (i) the attack technique is one of the three types of adversarial *spoofing* attacks (FGSM, PGD or GANBA) evaluated in this work; and (ii) the defense technique consists of applying adversarial training using the generated adversarial *spoofing* attacks of one of the three black-box methods (FGSM, PGD or GANBA) evaluated in Section 5.3 and generated with the system which employs the SENet50 PAD discriminator. On the other hand, the three last rows of Table 2 shows the performance metrics of the biometric system when it is attacked by one type of adversarial *spoofing* attack (FGSM, PGD or GANBA) and, at the same time, it is defended using the PAD discriminator trained with the proposed GANBA framework described in Section 3.2.

As shown in Table 2, the adversarial training defense technique, which uses the generated GANBA attacks, significantly outperforms the adversarial training defense technique, which uses either FGSM or PGD attacks, by more than 2% of absolute EER_{joint} in both the PA and LA scenarios. However, the proposed GANBA adversarial *spoofing* attacks are not effectively detected when using adversarial training with either the FGSM or PGD attacks, since the EER_{joint} is more than 34% in both the PA and LA scenarios. This result highlights the effectiveness of the proposed GANBA attacks. It is also very noticeable that adversarial training with the generated GANBA attacks is able to slightly improve the results of the baseline complete biometric system which is not exposed to any adversarial *spoofing* attacks. This can be due to the effect of data augmentation on the generated GANBA attacks which helps to detect even more original *spoofing* attacks of the ASVspoof 2019 evaluation dataset.

Nevertheless, the best defense technique is that of using the PAD discriminator trained with the proposed GANBA framework which outperforms the adversarial training defense technique in all cases. The usage of the resulting PAD system of the proposed GANBA framework is the best solution for defending the target biometric system. It is even able to significantly improve the results of the baseline biometric system which is not exposed to any adversarial *spoofing* attacks by 1.96% and 2.13% of absolute EER_{joint} in the PA and LA scenarios, respectively. This means that the trained PAD discriminator with the proposed GANBA framework is not only helpful for detecting adversarial *spoofing* attacks but also for detecting the original *spoofing* attacks of the ASVspoof 2019 test dataset.

Table 2. Results of the TDNN + PLDA/LCNN biometric system evaluated on the ASVspoof 2019 logical access (LA) and physical access (PA) test datasets in terms of EER_{PAD}(%), EER_{ASV}(%), EER_{joint}(%) and min-tDCF. The FGSM and PGD attacks employs their best attack configuration ($\epsilon = 1.0$). (*) The EER_{ASV} metric is evaluated considering both *bonafide* and *spoofing* utterances.

Adv. Attacks / Defenses		**Logical Access Attacks**				**Physical Access Attacks**			
Adv. Attack	**Adv. Defense**	**EER_{PAD} (%)**	**EER_{ASV} (%)**	**EER_{joint} (%)**	**min-tDCF**	**EER_{PAD} (%)**	**EER_{ASV} (%)**	**EER_{joint} (%)**	**min-tDCF**
No Processing	No Processing	5.85	30.58 *	19.87	0.1237	4.62	18.43 *	13.12	0.1221
PGD	Adv. Train FGSM	8.42	30.78 *	24.68	0.1532	7.50	19.87 *	16.42	0.1434
GANBA	Adv. Train FGSM	20.34	29.18*	40.12	0.3304	16.23	17.56 *	35.43	0.3012
FGSM	Adv. Train PGD	7.34	31.03 *	23.75	0.1476	6.73	19.02 *	15.78	0.1389
PGD	Adv. Train PGD	8.07	30.89 *	23.12	0.1502	7.82	18.07 *	16.19	0.1476
GANBA	Adv. Train PGD	17.65	29.87 *	38.39	0.3009	15.03	17.10 *	34.32	0.2893
FGSM	Adv. Train GANBA	5.14	30.51 *	18.95	0.1102	3.96	18.56 *	12.16	0.1084
PGD	Adv. Train GANBA	5.33	30.44 *	19.22	0.1204	4.22	18.31 *	12.34	0.1121
GANBA	Adv. Train GANBA	5.45	30.66 *	19.66	0.1222	4.37	18.72 *	12.48	0.1145
FGSM	PAD from GANBA	4.12	30.33 *	17.74	0.1011	3.21	18.22 *	11.16	0.0976
PGD	PAD from GANBA	4.46	30.74 *	18.15	0.1056	3.53	18.31 *	11.24	0.0996
GANBA	PAD from GANBA	4.75	30.62 *	18.53	0.1079	3.68	18.56 *	11.36	0.1015

6. Conclusions

In this paper, the robustness of state-of-the-art complete (ASV+PAD) voice biometric systems against adversarial *spoofing* attacks was studied. Furthermore, we proposed a novel generative adversarial network for biometric anti-spoofing (GANBA) capable of fooling the anti-spoofing system without being detected by the ASV system, i.e., without changing the speaker information of the utterance. Furthermore, we employed the generated attacks for defending the system by either applying adversarial training or using the resulting PAD discriminator of the proposed GANBA framework.

Experimental results showed that voice biometric systems are highly vulnerable to adversarial *spoofing* attacks in both physical and logical access scenarios. However, we showed that the biometric system can be effectively defended using the PAD discriminator of the proposed GANBA system. In fact, the proposed defense technique resulted in being more robust against both adversarial and original *spoofing* attacks. It is worth noting that the results presented here with the GANBA attacks can be directly compared with those of the adversarial biometric transformation network (ABTN) attacks in [15]. This comparison shows that although ABTN and GANBA follow the same strategy for attack generation, GANBA goes a step further in being trained as a generative adversarial network (GAN), and shows a higher capability for fooling the voice biometric system.

In the future, we would like to explore a cross-database evaluation of the proposed defense technique for voice biometric systems against adversarial *spoofing* attacks in order to study their generalization between different datasets [50].

Author Contributions: The individual contributions are provided as follows. Contributions: conceptualization, A.G.A., J.A.G.-L. and A.M.P.; methodology, A.G.A., J.A.G.-L. and A.M.P.; software, A.G.A.; validation, A.G.A., J.A.G.-L. and A.M.P.; formal analysis, A.G.A., J.A.G.-L. and A.M.P.; investigation, A.G.A., J.A.G.-L. and A.M.P.; resources, A.G.A., J.A.G.-L. and A.M.P.; data curation, A.G.A. and J.A.G.-L.; writing—original draft preparation, A.G.A., J.A.G.-L. and A.M.P.; writing—review and editing, A.G.A. and J.A.G.-L.; visualization, A.G.A.; supervision, J.A.G.-L. and A.M.P.; project administration, A.M.P.; funding acquisition, A.M.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades Proyecto PY20_00902, and by the project PID2019-104206GB-I00 funded by MCIN/ AEI /10.13039/501100011033.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki. Ethical review and approval were waived for this study, due to the usage of public datasets.

Informed Consent Statement: Informed consent was obtained from all subjects involved in this study when acquiring the public datasets employed in this work.

Data Availability Statement: The ASVspoof 2019 datasets were used in this study. They are publicly available at https://datashare.ed.ac.uk/handle/10283/3336 (accessed on 5 December 2021).

Acknowledgments: Alejandro Gomez-Alanis holds a FPU fellowship (FPU16/05490) from the Spanish Ministry of Education and Vocational Training. Jose A. Gonzalez-Lopez also holds a Juan de la Cierva-Incorporación fellowship (IJCI-2017-32926) from the Spanish Ministry of Science and Innovation. Furthermore, we acknowledge the support of Nvidia with the donation of a Titan X GPU.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jain, A.K.; Ross, A.; Pankanti, S. Biometrics: A Tool for Information Security. *IEEE Trans. Inf. Forensics Secur.* **2006**, *1*, 125–143. [CrossRef]
2. Gomez-Alanis, A.; Gonzalez-Lopez, J.A.; Peinado, A.M. A Kernel Density Estimation Based Loss Function and its Application to ASV-Spoofing Detection. *IEEE Access* **2020**, *8*, 108530–108543. [CrossRef]
3. Gomez-Alanis, A.; Peinado, A.M.; Gonzalez, J.A.; Gomez, A.M. A Gated Recurrent Convolutional Neural Network for Robust Spoofing Detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1985–1999. [CrossRef]

4. Wu, Z.; Leon, P.L.D.; Demiroglu, C.; Khodabakhsh, A.; King, S.; Ling, Z.H.; Saito, D.; Stewart, B.; Toda, T.; Wester, M.; Yamagishi, J. Anti-Spoofing for Text-Independent Speaker Verification: An Initial Database, Comparison of Countermeasures, and Human Performance. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 768–783. [CrossRef]
5. ISO. Presentation Attack Detection. Available online: https://www.iso.org/standard/67381.html (accessed on 28 December 2021).
6. Wu, Z.; Kinnunen, T.; Evans, N.W.D.; Yamagishi, J.; Hanilçi, C.; Sahidullah, M.; Sizov, A. ASVspoof 2015: The First Automatic Speaker Verification Spoofing and Countermeasures Challenge. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015; pp. 2037–2041.
7. Korshunov, P.; Marcel, S.; et al., H.M. Overview of BTAS 2016 speaker anti-spoofing competition. In Proceedings of the 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), Niagara Falls, NY, USA, 6–9 September 2016; pp. 1–6.
8. Kinnunen, T.; Sahidullah, M.; Delgado, H.; Todisco, M.; Evans, N.W.D.; Yamagishi, J.; Lee, K.A. The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection. In Proceedings of the Interspeech, 2017, Stockholm, Sweden, 20–24 August 2017; pp. 2–6.
9. Todisco, M.; Wang, X.; Vestman, V.; Sahidullah, M.; Delgado, H.; Nautsch, A.; Yamagishi, J.; Evans, N.; Kinnunen, T.; Lee, K. ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 1008–1012.
10. Yamagishi, J.; Wang, X.; Todisco, M.; Sahidullah, M.; Patino, J.; Nautsch, A.; Liu, X.; Lee, K.A.; Kinnunen, T.; Evans, N.; Delgado, H. ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. In Proceedings of the 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, Online, 16 September 2021; pp. 47–54.
11. Gomez-Alanis, A.; Gonzalez-Lopez, J.A.; Dubagunta, S.P.; Peinado, A.M.; Magimai.-Doss, M. On Joint Optimization of Automatic Speaker Verification and Anti-spoofing in the Embedding Space. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 1579–1593. [CrossRef]
12. Gomez-Alanis, A.; Peinado, A.M.; Gonzalez, J.A.; Gomez, A.M. A Deep Identity Representation for Noise Robust Spoofing Detection. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 676–680.
13. Gomez-Alanis, A.; Peinado, A.M.; Gonzalez, J.A.; Gomez, A.M. Performance evaluation of front- and back-end techniques for ASV spoofing detection systems based on deep features. In Proceedings of the Iberspeech 2018, Barcelona, Spain, 21–23 November 2018; pp. 45–49.
14. Gomez-Alanis, A.; Peinado, A.M.; Gonzalez, J.A.; Gomez, A.M. A Light Convolutional GRU-RNN Deep Feature Extractor for ASV Spoofing Detection. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 1068–1072.
15. Gomez-Alanis, A.; Peinado, A.M.; Gonzalez, J.A. Adversarial Transformation of Spoofing Attacks for Voice Biometrics. In Proceedings of the Iberspeech 2020, Valladolid, Spain, 24-25 March 2021; pp. 255–259.
16. Liu, S.; Wu, H.; yi Lee, H.; Meng, H. Adversarial Attacks on Spoofing Countermeasures of Automatic Speaker Verification. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 312–319.
17. Zhang, Y.; Jiang, Z.; Villalba, J.; Dehak, N. Black-box Attacks on Spoofing Countermeasures Using Transferability of Adversarial Examples. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 September 2020; pp. 4238–4242.
18. Ren, K.; Zheng, T.; Qin, Z.; Liu, X. Adversarial Attacks and Defenses in Deep Learning. *Engineering* **2020**, *6*, 346–360. [CrossRef]
19. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. In Proceedings of the International Conference on Learning Representations (ICLR), Banf, AB, Canada, 14–16 April 2014.
20. Hansen, J.H.L.; Hasan, T. Speaker Recognition by Machines and Humans: A Tutorial Review. *IEEE Signal Process. Mag.* **2015**, *32*, 74–99. [CrossRef]
21. Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5329–5333.
22. Prince, S.; Elder, J.H. Probabilistic Linear Discriminant Analysis for Inferences About Identity. In Proceedings of the IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8.
23. Ioffe, S. Probabilistic Linear Discriminant Analysis. In Proceedings of the European Conference on Computer Vision (ECCV), Graz, Austria, 7–13 May 2006; pp. 531–542.
24. Sizov, A.; Lee, K.A.; Kinnunen, T. Unifying Probabilistic Linear Discriminant Analysis Variants in Biometric Authentication. In Proceedings of the Structural and Syntactic Pattern Recognition, Joensuu, Finland, 20–22 August 2014; pp. 464–475.
25. Kenny, P. Bayesian Speaker Verification with Heavy-Tailed Priors. In Proceedings of the Odyssey 2010, Brno, Czech Republic, 28 June 28–1 July 2010.
26. Brümmer, N.; de Villiers, E. The Speaker Partitioning Problem. In Proceedings of the Odyssey 2010, Brno, Czech Republic, 28 June 28–1 July 2010.
27. Dempster, A.; Laird, N.; Rubin, D. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–38.
28. Lee, H.S.; Tso, Y.; Chang, Y.F.; Wang, H.M.; Jeng, S.K. Speaker Verification using Kernel-based Binary Classifiers with Binary Operation Derived Features. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 1660–1664.

29. Chingovska, I.; Anjos, A.; Marcel, S. Anti-spoofing in Action: Joint Operation with a Verification System. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 98–104.
30. van Leeuwen, D.A.; Brümmer, N. An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. In *Speaker Classification I: Fundamentals and Methods*; Springer: Heidelberg, Germany, 2007; pp. 330–353.
31. Doddington, G.R.; Przybocki, M.A.; Martin, A.F.; Reynolds, D.A. The NIST Speaker Recognition Evaluation—Overview, Methodology, Systems, Results, Perspective. *Speech Commun.* **2000**, *31*, 225–254. [CrossRef]
32. Sahidullah, M.; Kinnunen, T.; Hanilçi, C. A Comparison of Features for Synthetic Speech Detection. In Proceedings of the Interspeech 2015, Dresden, Germany, 6–10 September 2015; pp. 2087–2091.
33. Lavrentyeva, G.; Novoselov, S.; Malykh, E.; Kozlov, A.; Kudashev, O.; Shchemelinin, V. Audio Replay Attack Detection with Deep Learning Frameworks. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 82–86.
34. Todisco, M.; Delgado, H.; Evans, N.W.D. Constant-Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification. *Comput. Speech Lang.* **2017**, *45*, 516–535. [CrossRef]
35. Muckenhirn, H.; Magimai-Doss, M.; Marcel, S. End-to-End Convolutional Neural Network-Based Voice Presentation Attack Detection. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 335–341.
36. Kinnunen, T.; Delgado, H.; Evans, N.; Lee, K.A.; Vestman, V.; Nautsch, A.; Todisco, M.; Wang, X.; Sahidullah, M.; Yamagishi, J.; Reynolds, D.A. Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification: Fundamentals. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2195–2210. [CrossRef]
37. Leon, P.L.D.; Pucher, M.; Yamagishi, J.; Hernáez, I.; Saratxaga, I. Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 2280–2290. [CrossRef]
38. Sizov, A.; el Khoury, E.; Kinnunen, T.; Wu, Z.; Marcel, S. Joint Speaker Verification and Antispoofing in the i-Vector Space. *IEEE Trans. Inf. Forensics Secur.* **2015**, *10*, 821–832. [CrossRef]
39. Kinnunen, T.; Lee, K.; Delgado, H.; Evans, N.; Todisco, M.; Sahidullah, M.; Yamagishi, J.; Reynolds, D. t-DCF: A Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification. In Proceedings of the Odyssey 2018, Les Sables d'Olonne, France, 26–29 June 2018; pp. 312–319.
40. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
41. Deng, Y.; Karam, L.J. Universal Adversarial Attack Via Enhanced Projected Gradient Descent. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 1241–1245.
42. Pedroni, V.A. *Finite State Machines in Hardware: Theory and Design (with VHDL and SystemVerilog)*; The MIT Press: Cambridge, MA, USA, 2013.
43. Wang, X.; Yamagishi, J.; Todisco, M.; Delgado, H.; Nautsch, A.; Evans, N.; Ling, Z.H.; Becker, M.; Kinnunen, T.; Vestman, V.; et al. ASVspoof 2019: A large-scale public database of synthetized, converted and replayed speech. In *Computer Speech and Language*; Elsevier: Amsterdam, The Netherlands, 2020; p. 101114.
44. Nagrani, A.; Chung, J.S.; Zisserman, A. VoxCeleb: A Large-Scale Speaker Identification Dataset. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 2616–2620.
45. Kaldi ASR. SRE16 Xvector Model. Available online: http://kaldi-asr.org/models/m3 (accessed on 28 December 2021).
46. Lai, C.I.; Chen, N.; Villalba, J.; Dehak, N. ASSERT: Anti-Spoofing with Squeeze-Excitation and Residual Networks. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 1013–1017.
47. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
48. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; Devito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. In Proceedings of the Neural Information Processing Systems (NIPS), Long Beach, California, USA, 4–9 December 2017; pp. 1–4.
49. Uesato, J.; O'Donoghue, B.; Kohli, P.; van den Oord, A. Adversarial Risk and the Dangers of Evaluating Against Weak Attacks. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 5025–5034.
50. Korshunov, P.; Marcel, S. Cross-Database Evaluation of Audio-Based Spoofing Detection Systems. In Proceedings of the Interspeech 2017, San Francisco, California, USA, 8–12 September 2016; pp. 1705–1709.