Dictionary Attacks on Speaker Verification

Mirko Marras, *Member, IEEE*, Paweł Korus, *Member, IEEE*, Anubhav Jain, and Nasir Memon, *Fellow, IEEE*

***Abstract*— In this paper, we propose dictionary attacks against speaker verification-a novel attack vector that aims to match a large fraction of speaker population by chance. We introduce a generic formulation of the attack that can be used with various speech representations and threat models. The attacker uses adversarial optimization to maximize raw similarity of speaker embeddings between a seed speech sample and a proxy population. The resulting master voice successfully matches a non-trivial fraction of people in an unknown population. Adversarial waveforms obtained with our approach can match on average 69% of females and 38% of males enrolled in the target system at a strict decision threshold calibrated to yield false alarm rate of 1%. By using the attack with a black-box voice cloning system, we obtain master voices that are effective in the most challenging conditions and transferable between speaker encoders. We also show that, combined with multiple attempts, this attack opens even more to serious issues on the security of these systems.**

***Index Terms*— Authentication, biometrics (access control), speaker recognition, adversarial machine learning, impersonation attacks.**

I. Introduction

BIOMETRIC technologies constitute one of the most popular solutions to user authentication. They can offer high reliability and better user experience than classic password-based systems, especially on mobile devices [1]. Among the plethora of available modalities, the most commonly deployed verification systems look at faces [2], fingerprints [3], and speech [4] - all of which can be used in modern smartphones. In this study, we focus on speaker verification, a key component of voice assistants, which represent a rapidly growing human-computer interaction method popularized by smart speakers [5], [6].

Like other biometric modalities, speech remains susceptible to attacks [1] which target both speech recognition (e.g, by crafting hidden voice commands [7]) and speaker verification (e.g., impersonation via spoofing, re-play or voice synthesis/conversion [8], [9]). Speaker impersonation studied to date exclusively focuses on *targeted attacks*, which make two critical assumptions: (i) there is a specific single *victim* (i.e., a target identity whose voice the attacker tries to imitate) and (ii) a sample of the victim's voice is available (or needs to be obtained). While the required sample size varies, and tends to change depending on the attack method and authentication protocol (e.g., text-independent [10], [11], [12] or interactive challenge-response [13], [14]), the principle remains the same.

In this paper, we propose a novel attack vector against speaker verification systems: *untargeted dictionary attacks*. In contrast to targeted attacks, the goal is to match a non-trivial fraction of the user population by pure chance, without any knowledge of the victim's identity or voice. Such an attack could be leveraged for unlocking a phone found on the street or facilitating mass-scale voice commands to voice assistants in compromised home networks [15]. Our approach involves adversarial optimization of a novel attack objective and can be applied to arbitrary speech representations (e.g., waveforms, spectrograms, speaker embeddings), making it adaptable to different systems and verification protocols (e.g., text-dependent or independent). This attack opens up a novel threat against the voice modality.

The feasibility of dictionary attacks has recently been shown for the fingerprint [16], [17] and the face [18] modalities. The inspiration comes from *biometric menagerie* [19], a well-established principle of numerous biometrics to exhibit large variations of matching propensity across individuals. In particular, the most relevant group for our work is represented by people who tend to match others easily (*wolves*) and people highly susceptible to be matched (*lambs*). Dictionary attacks aim to exploit this phenomenon to generate *master biometric examples* that maximize the impersonation capability of generated samples. Combined with rapidly improving generative machine-learning models, e.g., generative adversarial networks [20] or variational auto-encoders [21], this attack may soon create the perfect storm for biometric authentication.

Our study makes the first step to formalize and extensively evaluate dictionary attacks against speaker verification systems. The main contributions of our work are listed below.

1) We propose a generic formulation of the attack based on adversarial optimization driven by raw similarity of speaker embeddings. The attack can be applied to various speech representation domains and threat models.
2) We evaluate the attack, comparing three speech representations and several speaker encoders, under white- and black-box settings, showing strong generalization to an unseen speaker population and (in some settings) non-trivial transferability to unseen encoders.

Manuscript received 24 April 2022; revised 19 October 2022; accepted 27 November 2022. Date of publication 15 December 2022; date of current version 23 December 2022. This work was supported by the National Science Foundation under Grant 1956200. The associate editor coordinating the review of this manuscript and approving it for publication was Ms. Elham Tabassi. *(Corresponding author: Mirko Marras.)*

Mirko Marras is with the Department of Mathematics and Computer Science, University of Cagliari, 09124 Cagliari, Italy (e-mail: mirko.marras@acm.org).

Paweł Korus was with the Tandon School of Engineering, New York University, Brooklyn, NY 10012 USA, and also with the Department of Telecommunications, AGH University of Science and Technology, 30-059 Kraków, Poland. He is now with Amazon, Seattle, WA 98109 USA (e-mail: pkorus@agh.edu.pl).

Anubhav Jain and Nasir Memon are with the Tandon School of Engineering, New York University, Brooklyn, NY 10012 USA (e-mail: aj3281@nyu.edu; memon@nyu.edu).

Digital Object Identifier 10.1109/TIFS.2022.3229583

3) We show that speaker verification systems are susceptible to this attack and that the effect varies across genders. In our experiments, an accidental intrinsic bias of speaker encoders made female speakers remarkably more vulnerable to the attack.

Compared to our prior study [22], we have revised and generalized the attack to enable seamless application to various speech representation domains. We also extended the evaluation to include several speaker encoders and various threat models. Our version in this paper leads to substantially better results and can be even used in challenging conditions, e.g., to evolve transferable master voices based on black-box access to a third-party voice cloning system with variable output.

II. Related Work

A. Speaker Modelling

Speaker recognition involves two main tasks [4]: *identification* aims to identify the speaker among a set of possible hypotheses; *verification* aims to confirm the identity of the claimed speaker and operates in an open-set regime based on a gallery of enrolled speech samples. Speaker modeling has recently been dominated by deep neural networks [23] (DNNs) which remarkably outperform classic solutions like GMM-UBM [24] or i-vector [25]. DNNs are typically pre-trained for the identification task, but are then adapted to open-set verification by discarding the classification head and extracting a compact intermediate representation, referred to as a *speaker embedding*. The embeddings are then compared between the query and enrolled samples to confirm the speaker's identity.

Speaker enrollment typically involves the collection of multiple speech samples, whose embeddings need to be combined. Some of the traditional methods (e.g., a PLDA model [4]) assume statistical independence, which is hard to achieve in practice. As a result, simpler scoring strategies are often preferred, e.g., averaging the embeddings or taking the one with maximum similarity. A recent study [26] showed that the average embedding often leads to superior performance, which makes it a popular choice [27], [28].

Countless model architectures have been proposed for speaker encoding. Some of the most prominent differences involve selection of the input acoustic representation, backbone network, and temporal pooling strategy. While directly using waveforms to learn a representation is possible [29], it is much more common to use a hand-crafted 2D representation (e.g., spectrograms or filterbanks). The latter enables adaptation of successful backbones from computer vision, e.g., VGG [30] or residual networks (ResNet) [31], [32], [33]. Dealing with the time dimension can rely on recurrence [34], pooling [33], [35] or specialized architectural designs. As an example, Time Delay Neural Networks (TDNNs) use a 1D convolution structure along the temporal axis and are adopted in the popular x-vector architecture [36], [37].

Usually, trainable pooling layers achieve better results than simple pooling operators, (e.g., average pooling [33] or statistical pooling [38]). Some of the most successful learned designs include the family of VLAD models. NetVLAD [39] assigns each frame-level descriptor to a cluster and computes residuals to encode the output features. Its variant GhostVLAD [39] improved performance by excluding some of the original NetVLAD clusters from the final concatenation, such that undesirable speech sections are down-weighted.

B. Adversarial Attacks in Speech Processing

Originally introduced in computer vision [20], adversarial attacks refer to genuine samples imperceptibly modified by tiny perturbations to fool classifiers with high chance. In the context of speech, this type of attack can be broadly categorized based on the targeted task, i.e., speech or speaker recognition. In the former, the goal is to embed carefully crafted perturbations to yield automatic transcription of a specific malicious phrase. In [40], the attacker uses inverse feature extraction to generate obfuscated audio played over-the-air, which allows for issuing hidden commands to voice assistants. Later, [41] proposed a white-box attack based on gradient optimization, leading to quasi-perceptible adversarial perturbations, finally improved using psychoacoustic modeling [42]. To avoid repeated optimization hindering real-time use, a recent work by [43] designed an algorithm to find a single universal perturbation, that can be added to any speech waveform to cause an error in transcription with high probability. Finally, [7] showed that adversarial commands can be also hidden in music. The authors used a surrogate model to create transferable adversarial examples that can achieve this goal.

Attacking speaker verification systems initially relied on spoofing and replay attacks. Susceptibility to adversarial examples has gained attention only recently. The goal is to craft an attack sample from a voice uttered by a seed speaker, so that it is misclassified as a different one (either specific or any), while still being recognized as the seed speaker by human listeners. In a white-box setting, the FGSM attack made it possible to generate adversarial examples with high success rate [9]. [44] constrained the perturbation based on a psychoacoustic masking threshold to obtain imperceptible samples. To obtain robustness against reverberation and noise, [45] proposed a gradient-based optimization that generates robust universal adversarial examples (though the attack was not tested over-the-air). Reference [46] used a gradient estimation algorithm (NES) in a black-box setting. While the study used a small dataset, the attack had a high success rate in a practical setting.

All of the existing attacks (including both spoofed and adversarial samples) are targeted, i.e., they aim to pass authentication as a specific individual. However, biometric systems exhibit large variations in matching propensity across individuals, which can be exploited to open a novel threat vector. Hence, the untargeted nature of the proposed dictionary attacks is fundamentally different from the untargeted nature of adversarial attacks on machine learning models. In this context, the latter would aim to prevent authentication as a particular person without specifying the desired target identity.

C. Dictionary Attacks in Biometrics

Dictionary attacks use prior knowledge about the expected success rate to triage brute-force authentication attempts. They

naturally apply to passwords, but until recently have not been considered for other authentication modalities. In biometrics, such attacks are qualitatively different from spoofing and do not require any knowledge about the victim (e.g., speech samples) [47]. This threat is enabled by large variation in matching propensity across individuals (biometric menagerie [19]) and further exacerbated by the usability-security trade-offs in mass deployments (e.g., partial finger impressions [16]).

The concept of dictionary attacks in biometrics was introduced only recently. The vulnerability was first demonstrated on fingerprints [16] and subsequently extended to faces [18]. Initially, an existing fingerprint with the highest impostor score was selected as a *master print* [16]. In the next iteration, synthetic master prints were created by first-order hill-climbing, initialized on the most promising real fingerprints from the first approach. However, local search algorithms may get stuck in local minima or take a long time to converge. Reference [17] used diversity-quality evolution to address this issue and a generative adversarial network (GAN) to parametrize the search space. The same approach was successful for faces [18].

So far, dictionary attacks have not been studied for speech. Our preliminary work [22] demonstrated that adversarial optimization of spectrograms in a white-box setting consistently increases impersonation rates in VGGVox [30]. The resulting adversarial samples could match, on average, 20% (10%) of female (male) speakers in an unseen population. In this paper, we generalize our attack and test it against multiple systems and diverse speech representations. We achieved substantially improved impersonation rates and demonstrate non-trivial transferability across speaker encoders.

III. Problem Statement and Attack Methods

In this section, we formally define the problem and provide a generic formulation of the attack.

A. Speaker Verification Pipeline

We consider a standard text-independent speaker verification pipeline where a fixed-length speaker embedding $\mathbf{f} \in \mathbb{R}^e$ is extracted from a speech waveform of variable length $\mathbf{w} \in [-1, 1]^*$ using a *speaker encoder* ($\mathcal{E}$). Verification for a given user u involves comparison of the speaker embedding $\mathbf{f}$ extracted from a presented test sample with a set of *enrolled embeddings* $F_u = \{\mathbf{f}_{\mathbf{u},\mathbf{i}} : \mathbf{u} = u\}$. Without loss of generality, we assume a fixed number of enrolled samples per person (n). The number of collected samples and their combination depend on an enrollment and scoring strategy. We discuss each step in detail below. Our notation is shown in Table I.

a) Speaker encoder: The speaker encoder $\mathcal{E}$ is typically a DNN trained on an *acoustic representation* $\mathcal{A}(\mathbf{w}) = \mathbf{A} \in \mathbb{R}^{k\times *}$ (e.g., spectrogram or filter banks):

$$[-1, 1]^* \ni \mathbf{w} \xrightarrow{\mathcal{A}} \mathbf{A} \in \mathbb{R}^{k\times *} \xrightarrow{\mathcal{E}} \mathbf{f} \in \mathbb{R}^e \tag{1}$$

The model is typically pre-trained for fully supervised closed-world speaker classification on a large corpus with thousands of speakers. Ultimately, the classification head is discarded and the preceding layer is used to extract the *speaker embedding*.

TABLE I
Key Symbols and Notation Summary

waveform	$\mathbf{w}$	$\in [-1, 1]^*$
acoustic representation	$\mathbf{A}$	$\mathcal{A}(\mathbf{w}) \in \mathbb{R}^{k\times *}$
speaker embedding	$\mathbf{f}$	$\mathcal{E}(\mathbf{A}) \in \mathbb{R}^e$
speech parametrization/generation	$\mathcal{G}(\mathbf{w}, \mathbf{v}\|\theta)$	
collection of speaker embeddings	F	$= \{\mathbf{f}_{\mathbf{u},\mathbf{i}}\} \in \mathbb{R}^{n\times m\times e}$
... for user u	F_u	$= \{\mathbf{f}_{\mathbf{u},\mathbf{i}} : \mathbf{u} = u\}$
... for population U_t	F^o	$= \cup_{u\in U_o} F_u$
optimization (proxy) population	U_o	
test population	U_t	
# enrolled users	m	$\|U_t\|$
# enrolled samples	n	
# presentation attempts	c	

b) Enrollment and scoring strategy: The system collects multiple speech samples of each user and stores the resulting speaker embeddings, i.e., the database is a collection $F = \{\mathbf{f}_{\mathbf{u},\mathbf{i}}\}$ where $\mathbf{u}$ denotes users and $\mathbf{i}$ indexes their successive samples (for simplicity, assume a constant number of samples n per user). During *verification*, the system returns a binary decision indicating whether a test sample matches a claimed identity u. The test waveform $\mathbf{w}$ is encoded analogously to enrollment and processed according to a verification rule:

$$v_{\rho,\tau}(\mathbf{w}, u|F) := v_{\rho,\tau}(\mathbf{f}, F_u) : \mathbb{R}^e \times \mathbb{R}^{n\times e} \to \{0, 1\} \tag{2}$$

which involves the choice of a scoring strategy ρ for combining multiple embeddings/scores and a threshold τ. We consider two popular policies [26]: (1) *any-n* scores similarity with each of the enrolled embeddings and takes the maximum one; (2) *avg-n* scores similarity to the average embedding. Formally:

$$v_{\text{any},\tau}(\mathbf{f}, F_u) = \text{any}\left(\{\mathbf{f} \circ \mathbf{f}_{u,\mathbf{i}} > \tau : \mathbf{i} = 1, \ldots, n\}\right) \tag{3}$$

$$v_{\text{avg},\tau}(\mathbf{f}, F_u) = \mathbf{f} \circ \mathbf{f}_u > \tau \tag{4}$$

$$= \mathbf{f} \circ \left(\frac{1}{n}\sum_{\mathbf{i}=1}^{n} \mathbf{f}_{u,\mathbf{i}}\right) > \tau \tag{5}$$

where operator $\mathbf{f}_1 \circ \mathbf{f}_2 \to \mathcal{R}$ denotes a similarity function (e.g., the cosine similarity or the inverse of the Euclidean distance).

B. Dictionary Attack Formulation

In contrast to classic speaker spoofing, the goal of dictionary attacks is to match a large fraction of an unknown population by pure chance. Formally, the goal of the attacker is to find a master voice sample $\mathbf{w}_*$ that maximizes false matching rate within some user population U:

$$\mathbf{w}_* = \underset{\mathbf{w}}{\text{argmax}} \underset{u\in U}{\mathbb{E}} \left[v_{p,\tau}(\mathbf{w}, u)\right] \tag{6}$$

This formulation assumes only a single presentation attempt and is referred to as a *master voice* (MV). However, many verification systems allow for several trials, each possibly using a different utterance. Hence, we distinguish a *maximum coverage master voice* sequence (MCMV) W_*:

$$W_* = \underset{(\mathbf{w}_1,\ldots,\mathbf{w}_c)}{\text{argmax}} \underset{u\in U}{\mathbb{E}} \left[v_{\rho,\tau}(\mathbf{w}_1, u) \vee \ldots \vee v_{\rho,\tau}(\mathbf{w}_c, u)\right] \tag{7}$$

This attack is more powerful, since each subsequent attempt can be optimized to target the remaining speaker embedding subspace. See Section III-F for more details.

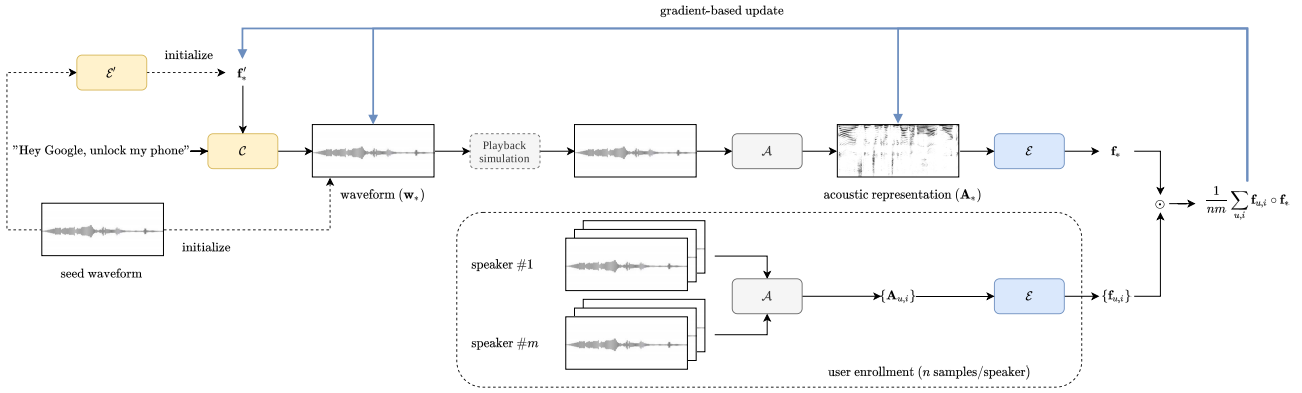


Fig. 1. Proposed adversarial optimization protocol for finding master voice samples: (1) a seed sample is used to initialize the attack; (2) a speaker embedding is computed and compared with a collection of enrolled embeddings from a proxy optimization population; (3) the gradient of the adopted similarity metric is computed and leveraged to update a chosen speech representation (e.g., waveform, spectrogram, or speaker embedding); depending on the adopted threat model either full gradient (white-box setting) or its approximation (black-box) is used.

C. Attack Implementation

The proposed attack is untargeted. Its goal is to maximize the *impersonation rate* (IR) in an unseen *test population* U_t. The IR is defined as the *expected fraction of user population that can be matched by an attack on an identity verification system*.[1] It can be considered for a single speech sample, or for a sequence of samples crafted for a multi-presentation attack (7). The distinction is clear from context. Formally, given a population U and the corresponding database of enrolled speaker embeddings F, the IR of a set of utterances W is estimated as:

$$\mathrm{IR}(W) = \frac{1}{|U|} \sum_{u \in U} \min\left(1, \sum_{\mathbf{w} \in W} v_{\rho,\tau}(\mathbf{w}, u|F)\right) \tag{8}$$

A practical implementation of our attack may use a proxy *optimization population* $\mathcal{U}_o$. Our implementation uses adversarial optimization driven by mean similarity of speaker embeddings in $\mathcal{U}_o$:

$$\mathbf{w}_* = \operatorname*{argmax}_{\mathbf{w}} \mathcal{S}(\mathbf{w}, F^o) \tag{9}$$

$$\mathbf{f} = \mathcal{E}(\mathcal{A}(\mathbf{w})) \tag{10}$$

$$\mathcal{S}(\mathbf{w}, F^o) = \frac{1}{nm} \sum_{u \in \mathcal{U}_o} \sum_{i=1}^{n} \mathbf{f} \circ \mathbf{f}_{u,i} \tag{11}$$

where m is the number of speakers and F^o denotes the gallery of speaker embeddings from the optimization population. As we will demonstrate in the experimental section, the attack transfers between different user populations. Due to observed distinct characteristics of male and female speech (and the resulting remarkable differences in their impersonation susceptibility [22]), we focus on attacking a single gender at a time. To avoid unnecessary complexity, we do not reflect this in our notation and simply remark that, unless stated otherwise, we assume and report results for each gender separately.

[1]Our definition of IR can be seen as a special case of spoof false acceptance rate (SFAR) commonly used in biometric literature and defined as *the number of times an active attack or an impostor is accepted as legitimate divided by the total number of attack or impostor attempts* [48].

We show a schematic illustration of the attack in Fig. 1. The process starts with a *seed* sample ($\mathbf{w}_0$) used for initialization of the attack. Then, the speaker encoder $\mathcal{E}$ extracts the embedding $\mathbf{f}_*$ for the optimized sample and its similarity to pre-computed embeddings from the optimization population is calculated as in (11). The gradient of the similarity score is then used to iteratively update the chosen representation of the speech sample. Let $\mathtt{t}$ denote the current step and T the total number of steps. The update process has the following general form:

$$\mathbf{v}^{\mathtt{t}+1} = \mathbf{v}^{\mathtt{t}} + \lambda \nabla_{\mathbf{v}} \mathcal{S}(\mathcal{G}(\mathbf{w}_0, \mathbf{v}^{\mathtt{t}}|\theta), F^o) \tag{12}$$

$$\mathbf{w}_* = \mathcal{G}(\mathbf{w}_0, \mathbf{v}^T|\theta) \tag{13}$$

where $\mathcal{G}(\mathbf{w}, \mathbf{v}|\theta)$ defines a speech representation/generation function driven by an adversarial *attack vector* $\mathbf{v}$ appropriate for that optimization domain. Our attack is generic and can work with various domains. We experimented with optimization of waveforms, spectrograms, and speaker embeddings (e.g., voice cloning). Each domain has its own peculiarities:

- *Waveform:* this attack aims to find an adversarial perturbation directly in the waveform domain. It starts with a seed sample and has a simple formulation that allows for straightforward inclusion of data augmentation, e.g., via playback simulation (Section III-E):

$$\mathcal{G}(\mathbf{w}, \mathbf{v}|\theta) = \mathbf{w} + \mathbf{v} \tag{14}$$

- *Spectrogram:* this attack aims to find an adversarial perturbation in the acoustic representation accepted by the speaker encoder as an input. It starts with a seed sample and requires invertibility of the representation to reconstruct the adversarial waveform:

$$\mathcal{G}(\mathbf{w}, \mathbf{v}|\theta) = \mathcal{A}^{-1}(\mathcal{A}(\mathbf{w}) + \mathbf{v}) \tag{15}$$

  Practical solutions would short-circuit inversion and feed the distorted acoustic representation to the encoder. Inversion can be performed once at the end. For spectrograms, we used the Griffin-Lim algorithm [49], but some acoustic representations may not be invertible.
- *Speaker embeddings:* this attack finds a voice that maximizes impersonation by optimizing a speaker embedding

$\mathbf{f}'$ used for conditioning a speech generation model $\mathcal{C}$:

$$\mathcal{C}(\mathbf{f}', \theta) = \mathbb{R}^{e'} \times \mathbb{T}^* \rightarrow \mathbf{w} \in [-1, 1]^* \tag{16}$$

where $\mathcal{C}(\mathbf{f}', \theta)$ denotes a voice cloning system able to generate any utterance (with content specified by a string of tokens $\theta \in \mathbb{T}^*$) spoken by any speaker (specified by a speaker embedding $\mathbf{f}'$ returned by a different encoder $\mathcal{E}'$) [50], [51]. The attack sample generation is:

$$\mathcal{G}(\mathbf{w}, \mathbf{v}|\theta) = \mathcal{C}(\mathbf{v}, \theta) \tag{17}$$

In our implementation, we used a seed sample to initialize the embedding, i.e., $v^0 = \mathcal{E}'(\mathbf{w}_0)$, but one may skip this step and simply sample a random one instead.

Following the same logic, the attack can be defined in other domains too, e.g., the latent space of a generative adversarial network [52], a (variational) auto-encoder [53], or the speaker embedding of a voice conversion system [54].

In practice, we implemented the attack using stochastic gradient descent. Processing the optimization population in batches is both necessary (due to constraints on available GPU memory) and advantageous to the speed of convergence (making more update steps requires fewer epochs). Depending on the size of the speaker encoder, we used batches of 64-256 items. We pre-shuffled the stored embeddings to diversify speaker identities within each batch. Choice of the update step λ and the number of epochs depend on the attack configuration. To speed up convergence and parameter choice, we use gradient normalization (with L_2 and L_∞ norms). The impact of these parameters is shown in detail in Section V-A.

D. White-Box Vs. Black-Box Attacks

Our attack can be carried out under both the *white-box* and *black-box* threat models. The white-box attack uses gradients computed by automatic differentiation in machine learning frameworks. This allows for fast and accurate optimization but is limited to known models operating in a fully differentiable pipeline. In contrast, the black-box attack uses surrogate gradients estimated by querying any model. This leads to a general attack applicable to all pipelines, but requires larger computational cost. We used the natural evolution strategy (NES) [55] to estimate the gradient based on a small number of queries with a Gaussian search distribution with antithetic sampling. Rather than maximizing an objective function directly, NES maximizes the expected value of the objective in the vicinity implicitly defined by a stochastic search distribution. This allows for gradient estimation in fewer queries than typical finite-difference methods. We leveraged NES as follows:

$$\nabla s(\mathbf{w}) \approx \nabla \mathop{\mathbb{E}}_{\mathbf{w}'}[s(\mathbf{w}')] \approx \frac{1}{2\,s\sigma} \sum_{i=1}^{s} \delta_i s(\mathbf{w} \pm \sigma \delta_i) \tag{18}$$

$$\delta_i \sim \mathcal{N}(0, \mathbf{I}) \tag{19}$$

$$s(\mathbf{w}) = S(\mathbf{w}, F^*) : F^* \sim F^o \tag{20}$$

where F^* is a batch sampled from the optimization population.

Such an attack requires fine-tuning of two additional hyperparameters (s, σ) but can be effective against various models. We discuss this in more detail in Sections V-C and V-E.

E. Playback Simulation

To assess (and improve) robustness to distortions, we include an optional *playback simulation* step. This step can be included both during evaluation and attack optimization. Let $\circledast$ denote 1D convolution and $(\mathbf{k}_s, \mathbf{k}_r, \mathbf{k}_m)$ denote the impulse responses of the speaker, room, and microphone, respectively. The waveform after playback can be computed as:

$$\mathbf{w}' = (((\mathbf{w} \circledast \mathbf{k}_s) + \mathbf{n}) \circledast \mathbf{k}_r) \circledast \mathbf{k}_m \tag{21}$$

$$\mathbf{n} \sim \mathcal{N}(0, \alpha \mathbf{I}) \tag{22}$$

To increase augmentation diversity, we randomize the simulation by sampling the AWGN strength $\sqrt{\alpha} \sim \mathcal{N}(0, 0.025)$ and choosing random kernel combinations from a small database with 4 speakers, 9 rooms and 7 microphones.

F. Multiple Presentation and Coverage Optimization

Due to their imprecise nature, biometric systems often allow for a few authentication attempts before falling back to a PIN or passphrase. This behavior can be exploited and the attacker can craft a sequence of diverse speech samples that maximize the overall success rate. Let c denote the number of allowed attempts. The attacker can simply generate a set of master voices $\{\mathbf{w}_\mathbf{c} : \mathbf{c} = 1, \ldots, c'\}$, optimized independently as in Eq. (6) based on $c' \gg c$ randomly chosen seed samples. Finally, the best c samples $(\mathbf{w}_{\mathbf{c}_1}, \ldots, \mathbf{w}_{\mathbf{c}_c})$ are chosen for the attack.

In our experiments, we simply reuse the *optimization* population U_o to assess viability of candidate samples. The attacker computes $\mathbf{B} = [b_{\mathbf{c},\mathbf{u}}]$, a binary *impersonation matrix* indicating matching success for the $\mathbf{c}$-th sample against user $u \in U_o$:

$$\mathbf{B}_{c' \times m} = \begin{bmatrix} v_{\rho,\tau}(\mathbf{w}_1, u_1) & \ldots & v_{\rho,\tau}(\mathbf{w}_1, u_m) \\ \ldots & \ldots & \ldots \\ v_{\rho,\tau}(\mathbf{w}_{c'}, u_1) & \ldots & q v_{\rho,\tau}(\mathbf{w}_{c'}, u_m) \end{bmatrix}$$

Aggregation along the user dimension yields expected IRs. We hence test two simple strategies:

- naive *independent selection* takes the top-c speech samples based on their IR on the entire optimization population, i.e., the i-th sample is simply:

$$\mathbf{w}_{\mathbf{c}_i} : \mathbf{c}_i = \mathop{\text{argmax}}_{\mathbf{c} \notin \{\mathbf{c}_1, \ldots, \mathbf{c}_{i-1}\}} \frac{1}{m} \sum_{u \in U_o} b_{\mathbf{c},u} \tag{23}$$

- *complementary selection* takes a single best sample step-by-step, each time maximizing the IR on the still *uncovered subset* of the optimization population, i.e., the i-th sample is chosen as:

$$\mathbf{w}_{\mathbf{c}_i} : \mathbf{c}_i = \mathop{\text{argmax}}_{\mathbf{c} \notin \{\mathbf{c}_0, \ldots, \mathbf{c}_{i-1}\}} \frac{1}{|U_i|} \sum_{u \in U_i} b_{\mathbf{c},u} \tag{24}$$

$$U_i = \begin{cases} U_o & \text{for } i = 1 \\ U_{i-1} \setminus \{\mathbf{u} : v_{\rho,\tau}(\mathbf{w}_{\mathbf{c}_{i-1}}, \mathbf{u})\} & \text{for } i > 1 \end{cases} \tag{25}$$

An interesting extension of this problem would be to jointly optimize all C samples. We leave this aspect for future work.

TABLE II
DATASET PARTITION

Dataset	Partition	Scope	$\|U\|$[1]	$\|A\|$[2]	$\|A\|/\|U\|$
VoxCeleb1-Dev	P1	SV Train	1,211	148,642	122
VoxCeleb1-Test	P2	SV Eval.	40	4,874	122
VoxCeleb2-Dev	P3a	MV Opt.	1,000	50,000	50
	P3b	MV Eval.	1,000	100,000	100
	P3c	SV Train	3,994	895,664	224

[1] $|U|$ refers to the number of included users in total
[2] $|A|$ refers to the number of included utterances in total

IV. EXPERIMENTAL SETUP

In this section, we show our experimental setup and the details of the used datasets, speaker encoders etc. We explain our model pre-training and calibration, and provide an exhaustive benchmark evaluation of the resulting speaker verification systems, including key aspects of their menagerie analysis.

A. Datasets

We used two public datasets in our work: VoxCeleb [30] and LibriSpeech [56]. VoxCeleb [30] is a large, state-of-the-art dataset of human speech composed of two parts: VoxCeleb1 (dev set: 1,221 speakers and 148,642 utterances; test set: 40 speakers and 4,874 utterances) and VoxCeleb2 (dev set: 5,994 speakers and 1,092,009 utterances; test set: 119 speakers and 36,237 utterances). Both parts are fairly gender-balanced (55% and 61% of male speakers, respectively) and feature speakers from various ethnicities, accents, and age groups. Original videos used for speech extraction were shot in a wide range of challenging environments, including red carpet interviews, outdoor stadiums, indoor studios, speeches given to large audiences, excerpts from professionally shot multimedia, and amateur footage shot on hand-held devices. Crucially, they represent challenging real-world conditions which vary in background chatter, room acoustics, overlapping speech, recording equipment quality, and surrounding noise.

We divided the VoxCeleb speakers into disjoint partitions (Table II). First, we sampled two gender-balanced subsets of 1,000 people for master voice optimization and testing (partitions P3a and P3b, respectively). The remaining 5,205 speakers (P1 and P3c) were used for speaker encoder training. For consistency with standard evaluation methodology, we used the VoxCeleb1 test partition (P2) for speaker encoder benchmarking and calibration. In contrast, LibriSpeech is a clean and transcribed dataset with high-quality recordings of English speakers reading excerpts from audio books in studio conditions, used only in our final voice cloning experiments.

In all experiments, we use single-channel 16-bit audio with 16 kHz sampling rate. The waveforms are normalized to [0,1] and standardized to be 2.58 second long, which is achieved by random cropping (mostly) or zero-padding (occasionally).

B. Speaker Encoders

We used speaker encoders based on various CNN backbones and acoustic representations, including adapted VGG (VGGVox [30]) and ResNet models (ResNet 50 [30] and Thin ResNet [39]) trained on spectrograms, and x-vector based on filter banks [38]. VGG and ResNet models were adapted from computer vision to spectrogram inputs by replacing the last fully-connected (FC) layer with two layers: a FC one with support in the frequency domain and average pooling with support on the time domain. X-vector [36] is a TDNN, which allows neurons to receive signals spanning multiple frames. Given a filter bank, the first five layers operate on speech frames, with a time context centered at the current frame. A pooling layer aggregates frame-level outputs and computes mean and standard deviation. Two FC layers aggregate statistics across the time dimension. We used a GhostVLAD pooling layer [39], with 10 clusters plus 2 ghost clusters, on all models.

C. Pre-Processing and Training

We trained our speaker encoders from scratch using samples from 5,205 speakers (partitions P1 and P3c). We randomly sampled segments from each of their utterances and standardized the inputs to 2-second clips (by cropping or padding, respectively). No voice activity detection or silence removal was applied. Spectrograms (filter banks) were generated in a sliding window fashion using a Hamming window of width 25ms and step 10ms. We used 512-point (Fast Fourier Transforms) FFTs yielding spectograms of size 257×200 and filter banks of size 24×300 (frequency$\times$temporal). Each acoustic representation was normalized by subtracting the mean and dividing by the standard deviation of all frequency components in a single time step. The model was trained for classification using Softmax and the Adam optimizer, with an initial learning rate of 0.001, decreased by a factor of 10 after every 10 epochs.

D. Speaker Verification Performance

The pre-trained speaker encoders are deployed in a speaker verification system (Section III-A) by stripping their classification heads. We performed a detailed assessment of open-set speaker verification performance that includes two key aspects:

- *Raw performance:* we test discriminability of speaker embeddings on the standard VoxCeleb1 test pairs (37,720 pairs; partition P2). Based on the collected cosine similarities, we find the ROC and derive common metrics, e.g., area under the curve (AUC) and equal error rate (EER).
- *Deployment performance:* we test performance that accounts for the enrollment and scoring strategy. We used a larger population of 1,000 people (partition P3a).

The resulting evaluation will be used for threshold calibration - to summarize the behavior we focus on thresholds corresponding to the EER and a 1% false acceptance rate (FAR-1).[2] The obtained results are summarized in Table III (extended version is reported in Table A.I and the corresponding ROC curves are shown in Fig. A.1). To enable comparison with related work and investigate the potential gap in typical deployments, we distinguish between evaluation on full audio clips (as performed in the literature) and short utterances only (as used in real deployments). For the latter, we randomly crop a 2.58 second segment. On full-length clips from the standard

[2]False Rejection Rate (FRR) and False Acceptance Rate (FAR) are often referred to as "miss" and "false alarm rates" in speaker verification literature.

TABLE III
SPEAKER ENCODER MODELS AND THEIR BENCHMARK PERFORMANCE

	AUC				EER				FRR @ FAR1%			
	R^1	R^2	Any	Avg	R^1	R^2	Any	Avg	R^1	R^2	Any	Avg
VGGVox	0.95	0.98	0.90	0.93	11.8	6.9	14.5	11.3	52.1	27.0	43.2	23.2
ResNet 50	0.96	0.98	0.92	0.94	9.9	5.2	14.0	10.8	43.7	19.9	37.6	18.6
Thin ResNet	0.97	0.98	0.92	0.94	9.1	5.6	14.7	11.3	37.3	18.5	39.3	20.3
X-Vector	0.96	0.97	0.91	0.93	10.9	8.2	16.0	12.5	40.2	28.2	44.8	29.2

R^1 standard VoxCeleb test pairs (no enrollment), tested on short (2.58 s) clips
R^2 standard VoxCeleb test pairs (no enrollment), tested on full-length clips

test set, our models reach EER of ≈5-8% - higher than the best reported results (≈2.5-5%), but reasonable given our substantially smaller training population.[3] On shorter clips, this deteriorates across all models down to ≈9-11%.

Enrollment of multiple samples and using a scoring strategy (tested on short clips only) can substantially improve performance, especially in the low FAR regime. We observed the best results with the *avg*-10 policy, which is consistent with earlier findings [26]. The *any*-10 policy was consistently inferior to the use of even a single speaker embedding.

E. Dictionary Attack Implementation Details

We rely on two disjoint populations for master voice optimization (partition P3a) and testing (P3b). Each population contains 1,000 speakers. We treat male and female speakers separately since their speech exhibits distinct properties and leads to differences in verification performance and vulnerability to impersonation attacks. Specifically, a menagerie analysis on the seed utterances included in partition P3a, whose details are reported in Fig. A.2, showed that women often have a higher average imposter score. This gender-wise difference is emphasized when we consider the impersonation rates achieved by the same seed voices against users from the two genders. Fig. A.3 shows that women tend to be impersonated more, even under the most secure setting (*avg*-10, raw *far*-1 threshold). VGGVox and x-vector are the least secure systems and exhibit the largest difference in the maximum IR between genders.

The optimization process starts with a seed sample. We randomly sampled 100 seeds for both male and female speakers from P3a. While this step could also possibly be exploited to further improve IRs, we opted for fully random selection to simplify the experiments and rely on a single set of seed voices regardless of the target model. The way seed voices are used differs among the attacks. When optimizing waveforms or acoustic representations, we start with the full content of a seed sample for the target gender. For other attacks, e.g., based on voice cloning, we used seed samples to initialize speaker embeddings that condition the generator. Based on the adopted representation and capabilities of the synthesis model, seed samples may not be needed.

During the attack, the speaker encoder operates in a configuration which compares the current attack sample with a batch of samples from the optimization population. We shuffle samples from various users to promote speaker diversity within each batch. After each batch, we normalize the gradients.[4] and apply the update. Using stochastic gradient descent is beneficial and leads to remarkably reduce optimization time (fewer passes over the entire population, compared with gradient accumulation). We used batches of 64-256 samples, based on the model size and GPU memory (single RTX 8000 GPU).

To monitor, we track IRs for a single scoring strategy after each epoch (*any*-10, raw *far*-1 threshold). At this stage, we stay with the adopted representation. If applicable, we return to the waveform domain (e.g., Griffin-Lim inversion [49]) after optimization ends. We then test speaker verification performance and compare IRs for seed-master voice samples. We initially compare various enrollment policies and decision thresholds, but then focus on one representative configuration.

V. EVALUATION OF THE PROPOSED ATTACK

In this section, we perform a detailed evaluation of the proposed attack. First, we compare two speech representation domains (waveforms and spectrograms) and investigate the impact of attack and speaker verification settings. The following experiments focus on a single system configuration (*avg*-10, raw *far*-1 threshold) and address playback simulation, threat models (white-box vs. black-box) and transferability. We then show efficacy of our attack in a challenging setup with black-box access to a voice cloning system able to generate master voices with arbitrary content. Finally, we consider coverage experiments with multiple presentation attempts.

A. Impact of Attack and Verification Settings

We first explore the impact of attack and speaker verification settings and measure the success (impersonation) rates. We target the VGGVox encoder and consider various scoring strategies and threshold settings. We consider representative attack variations including optimization in the spectrum and waveform domains and with different update steps, including both L_2 and L_∞ normalization and various step sizes λ (for the L_∞ variant with budget ϵ, we use steps size $\lambda = \frac{\epsilon}{10}$ to allow for more flexibility).

Fig. 2 shows how the female IR changes with successive epochs (passes over the entire population) for various step sizes (λ). Successive columns correspond to spectrum optimization with L_2 gradient normalization (1st column), and waveform optimization with L_2 (2nd) and L_∞ normalization (3rd). We can observe that our attack is highly effective - it substantially improves IRs across various settings and transfers well between user populations. For the monitored *any*-10 policy, the average impersonation rate on the unseen population increases from 7% to ≈66%. Convergence rates vary with step size, but the attack saturates at a comparable level. L_∞ tends to converge faster than L_2, but reaches lower success rates across all distortion levels (see column 4).

While at the time of the attack waveform and spectrogram optimization seem to reach similar IRs, the latter requires spectrogram inversion to yield an adversarial waveform

[3]Note that in contrast to the standard practice in speaker verification literature, we excluded 2,000 people from the training set (≈30% of the training population) for our master voice analysis. In our preliminary experiments we performed sanity checks on the entire population (≈ 7, 200 people) and were able to obtain EERs within 0.8% (percentage points) of the results in [30].

[4]To control the change in magnitude for the gradients and facilitate the selection of the learning rate, we divided the gradients by their L_2 norm.

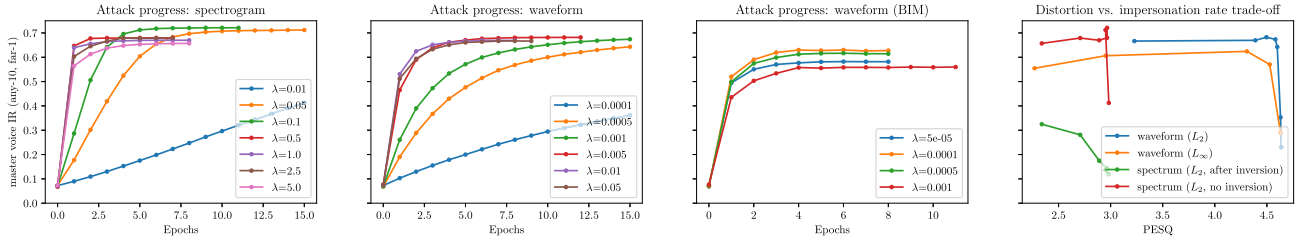


Fig. 2. Changes in IR (*any*-10 policy using raw *far-1* threshold) at successive optimization steps: (1st col.) spectrogram optimization; (2nd col.) waveform optimization with updates based on L_2-normalized gradient; (3rd col.) waveform optimization with a L_∞ constraint and binarized gradient; trade-off between attack success (impersonation) rate and distortion (PESQ).

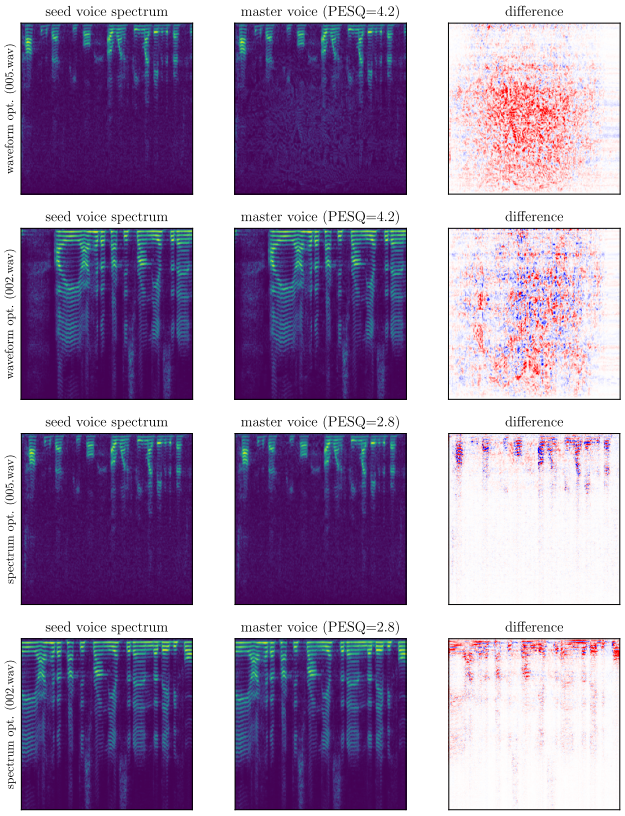


Fig. 3. Visualization of frequencies affected by the attack when optimizing the waveform (top 2 rows) and the spectrogram (bottom 2 rows); for the latter, we re-compute the spectrum from a reconstructed adversarial waveform.

(we used the Griffin-Lim algorithm). While the attack still works, it operates at an evidently reduced efficacy (down to $\approx 20-30\%$ in this experiment) and suffers from reduction of audio fidelity (see the gap in Perceptual Evaluation of Speech Quality PESQ [57] scores in column 4; higher scores for higher audio fidelity). We also visually compare the character of adversarial distortions in Fig. 3. Top rows depict 2 pairs of seed-master voice samples got with waveform optimization, and bottom rows depict 2 pairs of seed-master voice samples got with spectrum optimization. Waveform optimization affects a wider and higher range of frequencies.

In Table IV, we summarize IRs obtained for both female and male populations across several enrollment policies and decision thresholds. For clarity, we report only one set of

TABLE IV

AVERAGE IRS FOR SEED VOICES (SV) AND MASTER VOICES (MV) FOR VARIOUS SETTINGS SCORING STRATEGIES

	waveform optimization				**spectrum optimization**			
	female		**male**		**female**		**male**	
	SV	**MV**	**SV**	**MV**	**SV**	**MV**	**SV**	**MV**
any, τ : far1[1]	7.3	66.9	2.1	21.2	7.3	67.9	2.2	23.2
any, τ : far1	7.3	67.0	2.0	21.2	7.2	28.1	2.1	5.7
any, τ : eer	37.5	96.1	17.1	91.7	37.6	74.5	17.7	39.0
avg, τ : far1	6.9	84.7	1.6	63.3	6.7	37.7	1.7	10.7
avg, τ : far1[2]	2.4	69.5	0.4	38.0	2.5	19.4	0.5	3.5
avg, τ : eer	32.3	96.7	13.0	97.1	32.0	74.7	13.9	39.8

[1] optimization-time measurements without spectrogram inversion
[2] attack with good performance at a low distortion level

master voice samples which corresponds to a good trade-off between efficacy and audio fidelity. Our attack is effective regardless of system configuration and achieves non-trivial matching rates even in the most restrictive setting. At a *far*-1 threshold calibrated for the *avg*-10 policy, our master voice samples still impersonate 69% of females and 38% of males in a population unknown to the attacker.

Based on these results, in subsequent experiments we will restrict our attention to waveform-based attacks with L_2 gradient normalization and to speaker verification based on the *avg*-10 policy operating at a raw *far*-1 threshold.

B. Experiments With Playback Simulation

To test robustness of our attack to various distortions, we implemented playback simulation, which combines additive Gaussian noise with characteristics of a speaker, microphone, and surrounding environment (Section III-E). The simulation can be included both at testing and optimization time - assuming the representation domain precedes playback (e.g,. waveform or speaker embedding in synthesis systems). We therefore experiment with waveform optimization and assess the impact of the distortion at each of the mentioned stages.

In the following description, we use the terms *standard* and *augmented* optimization to indicate presence of playback. We randomly choose playback settings (noise strength, impulse response) for each batch. Apart from this, we follow the same experimental setup as before, i.e., we vary step size λ and measure IRs at various distortion levels. We show the obtained results in Fig. 4. The top row illustrates the trade-off between master voice IR and speech fidelity (PESQ).

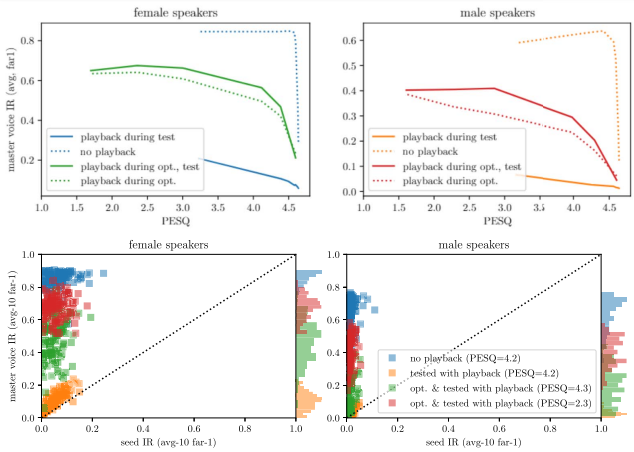


Fig. 4. Impact of playback simulation at the time of testing and optimization: (top) trade-off between IRs; (bottom) detailed scatter plot of (seed, master) IRs for selected configurations.

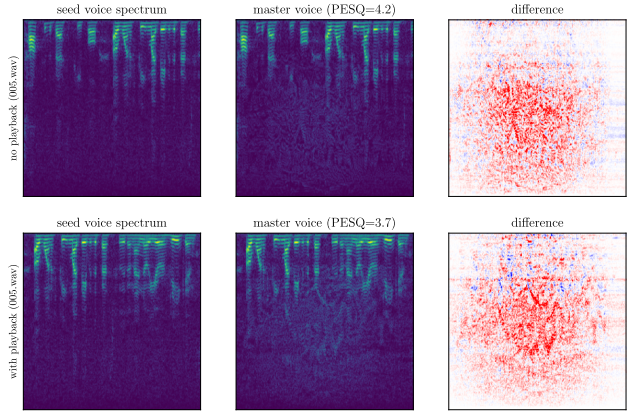


Fig. 5. Visualization of frequencies affected by waveform optimization with and without playback simulation.

We can observe that without augmented optimization, test-time playback (solid lines) renders adversarial waveforms nearly ineffective. The attack success rate can still increase somewhat, despite obvious saturation (or even rebound), during a standard test.

Augmentation leads to much more robust adversarial examples that achieve similar success rates with(out) playback (see solid vs. dotted lines of the same color). It leads to larger distortion, but does not obviously alter which frequencies are affected (Fig. 5). In both cases, the distortion tends to be the strongest in the middle of the sample. When audible, it sounds like a hissing modulated noise that does not interfere with the spoken content or the perceived identity of the speaker.

C. White-Box Vs Black-Box Attacks

Previous experiments relied on full gradients provided by automatic differentiation features in Tensorflow. While this leads to an effective attack, it is inflexible and often even impractical - due to either lack of knowledge or excessive implementation time to make everything fully differentiable. We hence switch to black-box optimization with gradients estimated by NES (Section III-D) from a similarity score

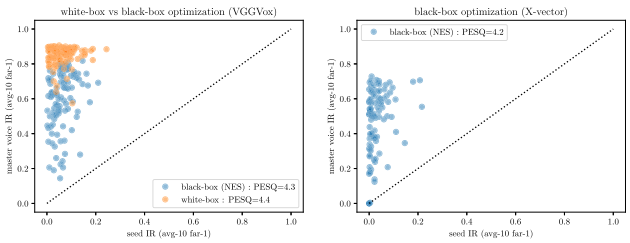


Fig. 6. Scatter plots of seed-master IRs for white-box optimization with accurate gradients vs. black-box optimization with NES-estimated gradients: (left) results for VGGVox at a similar distortion level; (right) results for x-vector where white-box optimization was not possible.

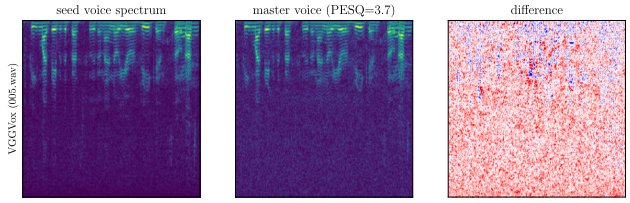


Fig. 7. Visualization of frequencies affected by waveform optimization in the black-box attack based on NES.

returned by the speaker verification system. Based on preliminary experiments on a small grid of feasible values, we set NES parameters to $s = 100$ samples and $\sigma = 0.001$. Due to large increase in computational requirements, we use a single step size $\lambda = 0.01$ and limit the number of epochs[5] to 10.

Fig. 6 compares white-box and black-box optimization for VGGVox (left) and shows black-box results for x-vector which uses non-differentiable filter-banks as its acoustic representation (right). In both cases, our black-box attack reaches IRs of 47% for x-vector and 59% for VGGVox (white-box attack at a comparable distortion reached 85%). Compared to white-box optimization, the black-box attack uniformly affected all frequencies at all times (Fig. 7).

D. Experiments With Transferability

We then test master voice transferability between speaker encoders: ResNet 50, Thin ResNet, VGGVox (all based on spectrograms) and x-vector (based on filter banks). We used waveform optimization in the white-box setting and fall back to black-box NES updates for x-vector. We test all combinations of playback simulation (optimization and testing time).

We collected results in Table V (female speakers). In general, waveform optimization does not lead to transferable master voices. The obtained adversarial speech relies on carefully crafted noise (see Fig. 3, 5, 7) and not on changes in speaker characteristics. Playback simulation in augmented training does provide a small but consistent improvement in transferability, but insufficient for an effective attack.

[5]NES relies on many independent function calls, trivially parallel, and could be optimized. Our naive sequential implementation on one RTX 8000 GPU required approx. 10 minutes per epoch (with $s = 100$ function samples). The corresponding white-box optimization takes approx. 10 seconds per epoch.

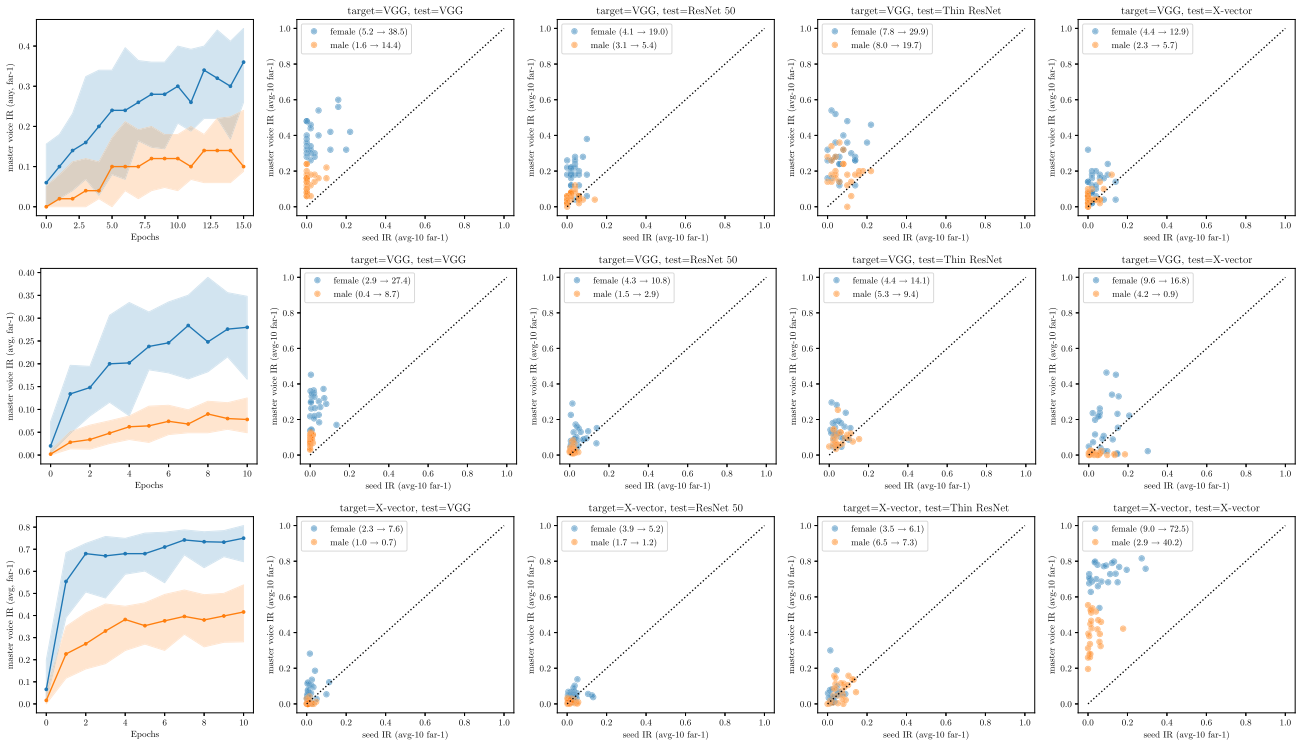


Fig. 8. Optimization progress and final impersonation rates of seed-master voice pairs obtained with voice cloning: (top) optimization targeting the VGGVox encoder on the LibriSpeech dataset; (middle) optimization targeting VGGVox on the VoxCeleb dataset; (bottom) optimization targeting x-vector on the VoxCeleb dataset. Targeting VGGVox tends to transfer to other models with best results observed for female speakers and other spectrogram-based encoders (ResNets). Targeting x-vector yielded stronger impersonation capabilities but poor transferability.

TABLE V

TRANSFERABILITY OF MASTER VOICE OBTAINED WITH WAVEFORM OPTIMIZATION TARGETING DIFFERENT SPEAKER ENCODERS

Target[2] \ Test[1]	tested w/o playback				tested w/ playback			
	R50	TR	VGG	X	R50	TR	VGG	X
standard optimization								
MV : ResNet 50	**35.7**	4.4	4.5	4.1	**4.8**	2.5	3.7	0.4
MV : Thin ResNet	3.2	**68.6**	5.2	4.4	2.8	**7.1**	4.1	0.4
MV : VGG	2.7	5.5	**89.6**	4.7	2.9	2.6	**9.3**	0.4
MV : X-vector[3]	5.1	9.0	7.4	**73.3**	5.3	4.8	5.3	**1.6**
SV : seed voice	2.5	4.6	4.7	3.5	2.8	2.4	3.8	0.5
augmented optimization								
MV : ResNet 50	**26.8**	4.9	6.1	4.2	**33.0**	3.0	5.3	**0.4**
MV : Thin ResNet	3.3	**43.3**	7.3	**4.9**	3.3	**51.9**	6.5	0.4
MV : VGG	3.4	6.3	**36.1**	5.1	3.4	3.5	**39.2**	0.5
SV : seed voice	2.5	4.5	4.8	3.6	2.9	2.3	4.1	0.4

[1] Abbrev.: (R50) ResNet 50; (TR) Thin ResNet; (VGG) VGGVox; (X) X-Vector
[2] Master voice examples were optimized with $\lambda = 0.01$
[3] uses black-box optimization

E. Experiments With Voice Cloning

Optimization in the waveform domain leads to highly effective adversarial speech samples (reaching even up to 85% IR) that can be made robust to various distortions via playback simulation. However, the optimization learns to embed carefully crafted noise that does not change the content or speaker identity and generally does not transfer between encoder architectures. In this section, we take advantage of the flexibility of our attack and experiment with a more compact, disentangled representation - we investigate optimization of the speaker embedding in a complex voice cloning system.

We used an open source system [51], [58] that generates speech based on a text prompt and a 256-d speaker embedding. The system uses Tacotron [59] for waveform synthesis, WaveRNN [60] as a vocoder, and an LSTM-based encoder [34]. All models are implemented in PyTorch, and we integrated the system with our Tensorflow-based attack framework via a simple black-box API that exposes two functions:

```
- get_speaker_embedding(speech_sample)
- generate_speech(text, speaker_embedding,
 max_len)
```

The generated output is stochastic and exhibits variations in sound and length of the waveforms. As a result, it represents a realistic and challenging attack scenario.

We fixed the text prompt to *"The assistant is triggered by saying hey google"* and use NES to evolve the speaker embedding, initialized from a seed voice by the black box speaker encoder $\mathcal{E}'$. Based on preliminary experiments, we set NES parameters to $s = 50$, $\sigma = 0.025$ and step size to $\lambda = 0.1$. We also clip[6] the embedding to stay within the expected domain of $[0, 1]^{256}$ and normalize the length of the output waveforms to 2.58 seconds. The cloning system was

[6]Clipping was performed to avoid deviations from the domain of seed vectors of the generative model. Departure from the commonly used n-dimensional hypercube or high-density regions of a standard multivariate Gaussian tend to introduce artifacts or break the synthesis entirely.

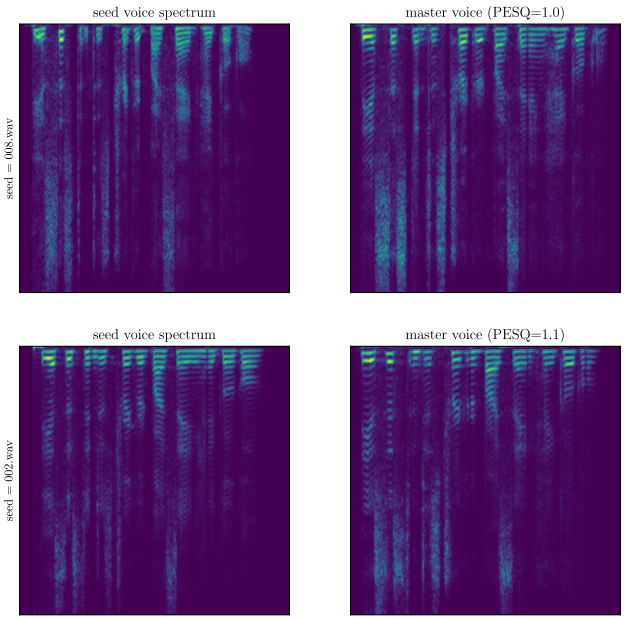


Fig. 9. Spectrograms of two pairs of seed and master voice samples obtained with voice cloning; the attack adapts the speaker characteristics and does not in result in tailored adversarial noise.

trained using LibriSpeech [61], so we conduct our experiments on this dataset as well. On VoxCeleb, we used the same setup as before. On LibriSpeech, we used a popular subset *train-clean-100* with 250 speakers which we split into two disjoint populations (optimization and testing) with 100 people, each with equal balance between genders. We randomly chose 10 samples per speaker both for optimization and for enrollment into the verification system. Due to much larger computational footprint, we repeat the attack based on 25 seed samples and run the attack for 15 (or 10) epochs for the LibriSpeech (VoxCeleb) datasets, respectively. We target the VGGVox and x-vector encoders and assess transferability.

We show the obtained results in Fig. 8. The left column shows attack progress for male and female speakers along with top and bottom percentiles (90-th and 10-th, respectively) of the observed impersonation rates (on the unseen test population). Despite the randomness of the generation and large variations in numbers, our attack consistently increases IRs for both male and female speakers, although the effect is substantially stronger for the latter. We compare the initial and final IRs using scatter plots (columns 2-5) for all considered speaker encoders. Each row corresponds to one targeted model (VGGVox or x-vector) and successive columns correspond to different test models and demonstrate transferability of the obtained samples. On the small LibriSpeech dataset (1st row) master voices optimized using VGGVox successfully transferred across all encoders. Again, the effect depends on the gender and tends to be much stronger for female speakers. On VoxCeleb the results are similar with the exception of male speakers tested on x-vector (which surprisingly has a strong negative effect). Targeting x-vector yielded much more effective, but generally non-transferable master voice samples - although for female speakers a weak effect seems to exist. We summarize the average impersonation rates in Table VI.

TABLE VI
TRANSFERABILITY OF MASTER VOICES OBTAINED WITH VOICE-CLONING

	seed voice				master voice			
	R50	**TR**	**VGG**	**X**	**R50**	**TR**	**VGG**	**X**
(LibriSpeech) far-1 calibrated on raw embedding similarity								
VGG (female)	4.1	7.8	5.2	4.4	19.0	29.9	**38.5**	12.9
VGG (male)	3.1	8.0	1.6	2.3	5.4	19.7	**14.4**	5.7
(VoxCeleb) far-1 calibrated on raw embedding similarity								
VGG (female)	4.3	4.4	2.9	9.6	10.8	14.1	**27.4**	16.8
VGG (male)	1.5	5.3	0.4	4.2	2.9	9.4	**8.7**	0.9
(VoxCeleb) far-1 calibrated on raw embedding similarity								
X-vector (female)	3.9	3.5	2.3	9.0	5.2	6.1	7.6	**72.5**
X-vector (male)	1.7	6.5	1.0	2.9	1.2	7.3	0.7	**40.2**

In contrast to waveform optimization, the attack does not result in obvious adversarial artifacts and appears to adapt the speaker characteristics (see Fig. 9). This may explain improved transferability between speaker encoders (Table VI and Fig. 8) and appears to match cross-system biometric menagerie evaluation. We assessed correlations of impersonation rates between all systems as a further validation (see Fig. A.4).

F. Experiments With Multiple Presentation Attempts

We finally evaluated seed and master voices in a setting where the speaker verification system allows users to do more than one attempt (we considered $c = 5$ allowed attempts). We tested two simple strategies (see Section III-F): naive independent selection (*ind*) and complementary selection (*comp*). To obtain more stable results, a strategy was repeated 100 times, each on a different subset of the seed/master voice population composed by 75% of randomly sampled users (results were averaged). We compared the two strategies against a random selection (*rand*). Due to their high impersonation power and transferability, we focus on master voice examples optimized (with playback) for the speaker embedding in a voice cloning system, targeting VGGVox (see Section V-E).

We collected the results for VGGVox in Fig. 10, for each gender separately (first two rows) and for a setting where we assume the attacker does not know the victim's gender (third row). In general, a complementary selection on master voices leads to the highest overall and cross-attempt IRs, except for x-vector. The adversarial speech relies on carefully crafted perturbations targeting a CNN-like architecture (VGGVox), and those perturbations might not be comparably effective on a different type of encoder architecture (x-vector is a TDNN based on filter banks; the others are CNN based on spectrograms). This is confirmed also by a transferability analysis in Fig. 14, which shows that FARs and IRs for seed utterances tend to not transfer between CNN-like encoders and x-vector. The explored selections obtain substantial gains for the male speakers, often doubling the IR of the best seed setting at that attempt. These two strategies also allow to improve transferability against Thin ResNet and ResNet 50. Similar observations were made while targeting x-vector during optimization (Fig. A.5).

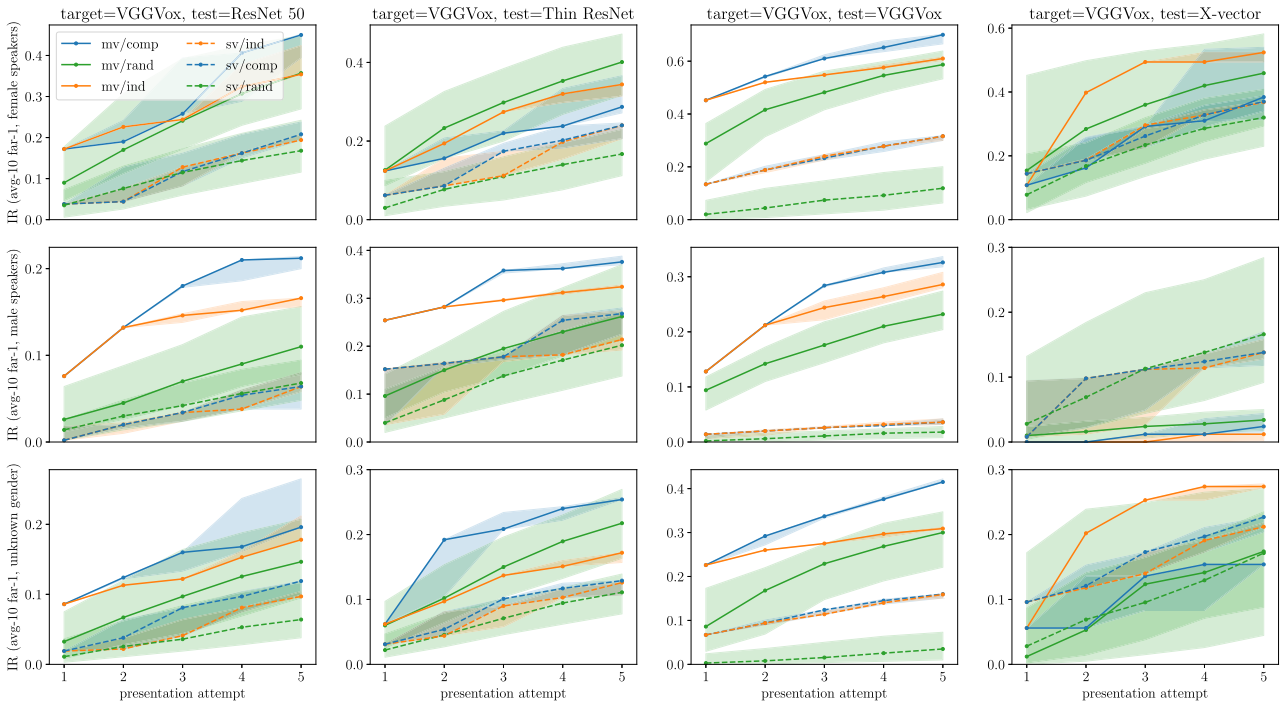


Fig. 10. Impersonation rates of seed and master voices under multiple presentation attempts ($n = 5$) in the black-box attack based on NES against the VGGVox speaker encoder, under an *avg*-10 policy with raw *far-1* threshold. (top) female gender, (middle) male gender, (bottom) unknown gender.

VI. Conclusion and Future Work

In this paper, we proposed dictionary attacks against speaker verification, a novel attack vector which aims to match a large fraction of user population by pure chance. In contrast to well known spoofing attacks that target one specific individual, our attack aims to exploit the biometric menagerie property - an inherent diversity in matching propensity/susceptibility across different people. Our approach is general and can be applied in various domains, including waveforms, acoustic representations, or even speaker embeddings in voice synthesis systems. We tested several different speaker encoder architectures and considered both white-box and black-box threat models.

We performed the first comprehensive evaluation of dictionary attacks against deep learning based speaker verification systems. The key conclusions from our work are as follows:

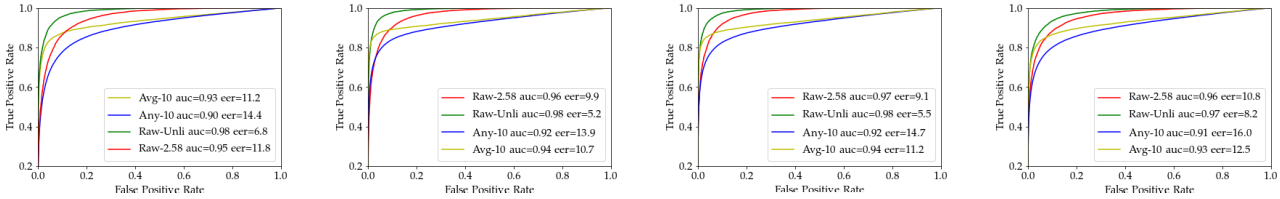
1) Speech appears to be susceptible to dictionary attacks. We were able to consistently and substantially increase IRs for all considered speaker encoders. Even for the most restrictive threshold (*far*-1 calibrated for the *avg*-10 scoring strategy), we were able to craft adversarial waveforms matching 69% of females and 38% of males in a population of 1,000 people (Table IV).
2) Susceptibility to the attack can vary remarkably across genders. We consistently observed much larger IRs for female speakers. The cause of this discrepancy is not clear. Our training set was only slightly imbalanced (64% of male speakers) and recent studies found only weak impact of gender balance on the overall error rates even for more unbalanced settings [62]. Further investigation of this aspect is needed.
3) Adversarial optimization driven by raw embedding similarity on a proxy population is a simple and effective attack strategy (e.g., it does not depend on configuration details, such as enrollment policy or decision threshold). The attack works well across speech representations (waveform, spectrogram, speaker embedding) in white- and black-box threat models. In a challenging scenario, our black box attack based on NES was highly effective even when targeting a complex black-box voice cloning system with highly variable output (Section V-E).
4) Our attack transfers across populations but not necessarily across genders. No notable differences in IRs between test and optimization populations were observed. Male and female speech tend to have different characteristics, and targeting both appears to be ineffective. We got best results when seed and target genders match.
5) Choice of speech representation has crucial impact on the attack. Optimization in the waveform and spectrogram domains leads to adversarial samples with crafted noise. Despite being very effective against the targeted model, it does not transfer between encoder architectures. Optimization of the speaker embedding in voice cloning led to a less effective but transferable attack.

Our results show that dictionary attacks could be a serious threat to speaker verification. We suspect there are two main factors at play. First, the speaker embedding space is likely not distributed uniformly. Regions of high and low density manifest themselves as differences in matching propensity/susceptibility which are characterized via the biometric menagerie. Our attack can take advantage of modern optimization methods and generative models to find speech properties that exploit this property. The higher transferability obtained through voice cloning suggests the existence of high-level

TABLE A.I
EXTENDED BENCHMARK PERFORMANCE FOR SPEAKER ENCODERS

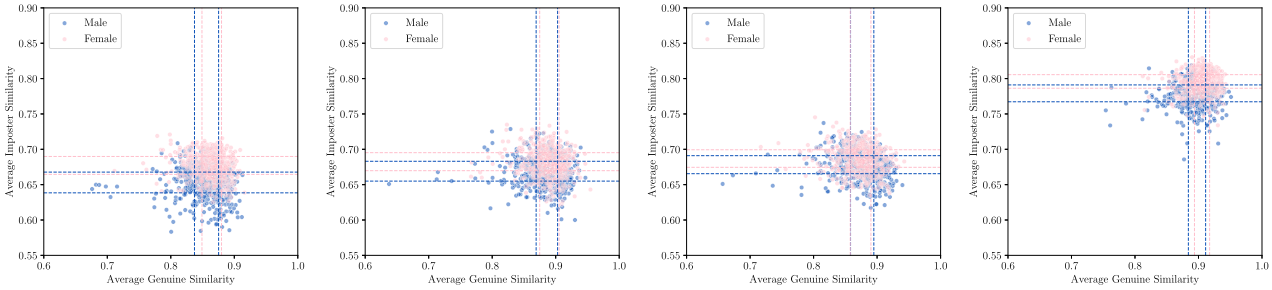
	$\mathcal{A}$ (k)	AUC				EER				EER Threshold				FRR @ FAR1%				FAR Threshold			
		Raw[1]	Raw[2]	Any-10	Avg-10	Raw[1]	Raw[2]	Any-10	Avg-10	Raw[1]	Raw[2]	Any-10	Avg-10	Raw[1]	Raw[2]	Any-10	Avg-10	Raw[1]	Raw[2]	Any-10	Avg-10
VggVox	S (256)	0.95	0.98	0.90	0.93	11.81	6.87	14.47	11.28	0.717	0.768	0.716	0.775	52.12	26.99	43.21	23.25	0.806	0.834	0.824	0.859
ResNet 50	S (256)	0.96	0.98	0.92	0.94	9.96	5.21	13.98	10.79	0.739	0.774	0.723	0.773	43.72	19.92	37.61	18.61	0.821	0.834	0.824	0.852
Thin ResNet	S (256)	0.97	0.98	0.92	0.94	9.11	5.56	14.75	11.28	0.738	0.769	0.715	0.775	37.33	18.47	39.26	20.28	0.802	0.815	0.807	0.844
XVector	F (24)	0.96	0.97	0.91	0.93	10.88	8.24	16.01	12.54	0.807	0.842	0.806	0.842	40.19	28.25	44.78	29.21	0.854	0.881	0.868	0.891

[0] acoustic representations (of size k): (S) spectrogram; (F) filter banks; [1] tested 2.58 seconds; [2] tested on full-length



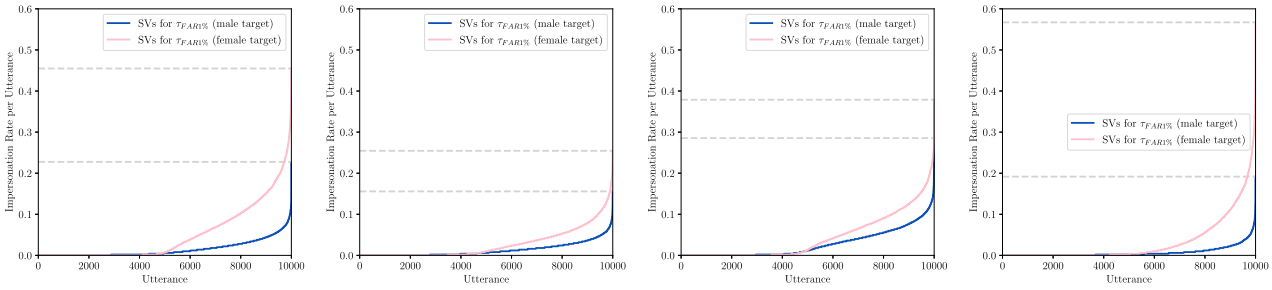
(a) VGGVox (b) ResNet 50 (c) Thin ResNet (d) X-vector

Fig. A.1. Receiver operation characteristics of the considered speaker encoders. We include both raw discriminability of the embeddings and final performance accounting for the enrollment and verification policy. It can be observed that the *avg* policy leads to a more secure system in the low FAR regime.



(a) VGGVox (b) ResNet 50 (c) Thin ResNet (d) X-vector

Fig. A.2. For each speaker encoder, a menagerie analysis plot. Each point in a plot represents a user, defined by their own average imposter score (when matched with other people) and the average genuine score (when matched with others of their own examples). For both the x- and y-axis, the two dashed lines indicate the 25% and 75% percentile. It should be noted that female users tend to have a higher average imposter score. The range of similarity scores is small, therefore even small differences in these average scores can determine highly important differences in impersonation rates between the genders. An ideal speaker recognition system should locate users at the lower right corner. Ideal impersonators would be located in the top part of the plot.

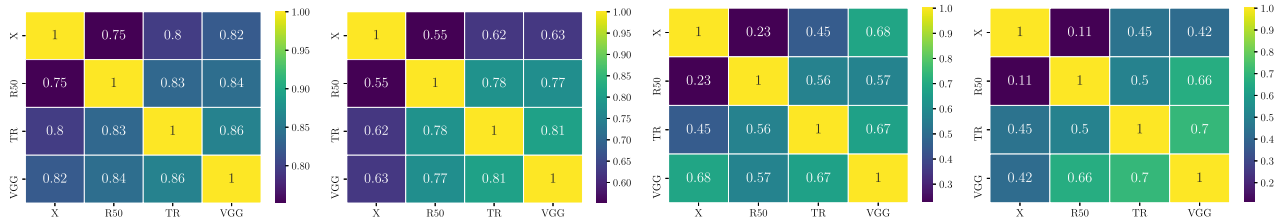


(a) VGGVox (b) ResNet 50 (c) Thin ResNet (d) X-vector

Fig. A.3. For each speaker encoder, a ranking of all seed utterances according to their Impersonation Rate (IRs) against male (blue) and female (pink) users, under *avg*-10 enrolment/verification policy with a raw *far*-1 threshold. Utterances are sorted by left to right based on an increasing IR. The higher the IRs, the more the utterance tends to match users. It should be noted that female users tend to be impersonated more. VGGVox and x-vector exhibit the larger difference in the maximum IR between genders and represent the least secure system.

master voice characteristics. Secondly, it appears that speaker encoders lack adversarial robustness and allow for finding noise-like perturbations that can maximize similarity even further.

That being said, our work has several limitations and should be seen as the first step in this direction. We designed our study to include both various speaker encoders and acoustic representations and to evaluate them fairly under the same conditions. Nevertheless, the state-of-the-art in both speaker verification and speech synthesis is moving quickly and more work will be needed to consider both classic approaches (e.g., GMM-UBM [24] or i-vector [25]) and emerging



(a) Cross-system FAR Female (b) Cross-system FAR Male (c) Cross-system IR Female (d) Cross-system IR Male

Fig. A.4. On the left (a) and (b), Spearman correlation respectively between the female and male False Acceptance Rates (FARs) raised by a given seed utterance between two speaker encoders, under an *avg*-10 enrolment and verification policy with a raw *far*-1 threshold. The higher the correlation, the more the utterances tend to have a high impersonation rate on both encoders. On the right (c) and (d), Spearman correlation respectively between the female and male Impersonation Rates (IRs) experienced by the same user between two speaker encoders, under an *avg*-10 enrolment and verification policy with a raw *far*-1 threshold. The higher the correlation, the more the same users end up being impersonated consistently between encoders. It should be noted that FARs and IRs tend to not transfer between CNN-like encoders and x-vector.

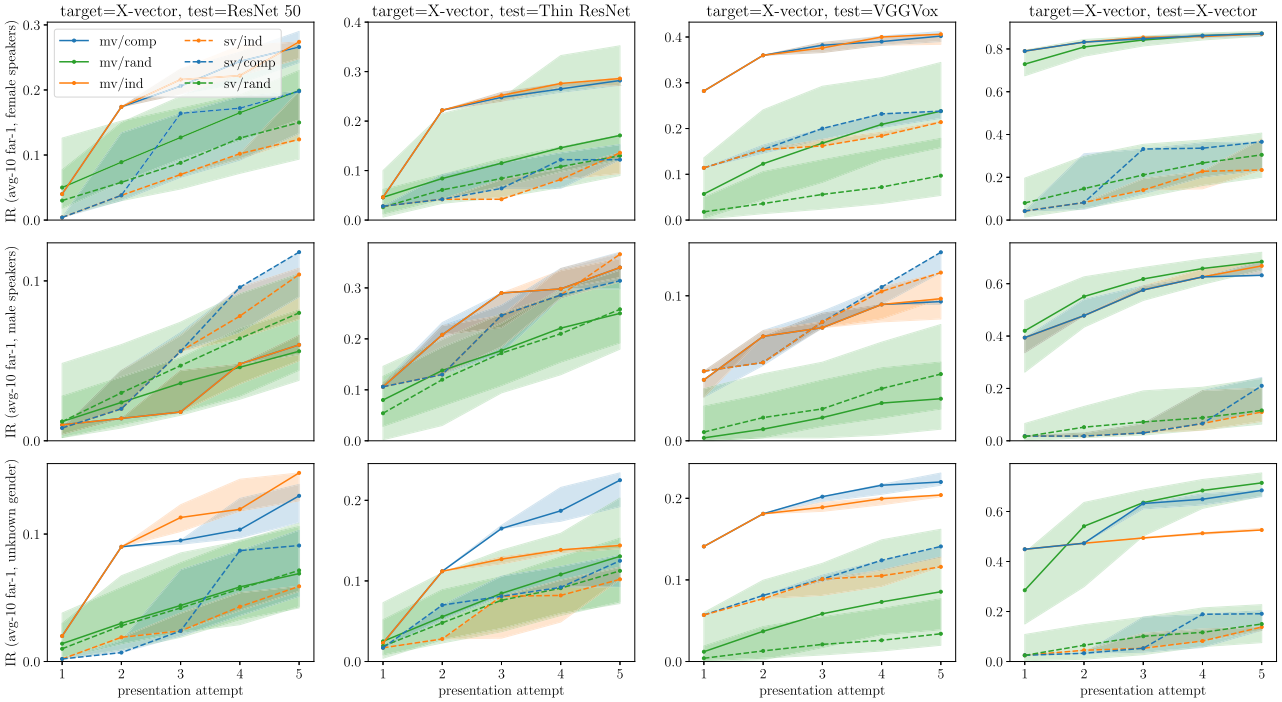


Fig. A.5. Impersonation rates of seed and master voices under multiple presentation attempts (n=5) in the black-box attack based on NES against the X-vector speaker verification system, under an *avg*-10 enrolment/verification policy with a raw *far*-1 threshold. It should be noted that master voice samples appear to generalize well across populations within X-vector and to transfer well to VGGVox. Transferability performance for the other systems is lower than for the former, though the master voice samples still lead to substantially higher impersonation rates than seed voice examples.

architectures (e.g., TDNN [63] or s-vectors based on transformers [64]).

The second main limitation of our current attack is the need for validation in a real over-the-air setting. So far, we relied on playback simulation which reveals that the signal distortion introduced by the channel can most likely be dealt with by means of augmented training. However, attacking real deployments will also need to address other factors, e.g., temporal shifts stemming from unknown start and duration of the sample. Another limitation is that both the feasibility of the attack and applicability of various potential countermeasures will come down to the root cause of the vulnerability. While *noise-based adversarial samples* will be difficult to use in practice (e.g., due to unknown model architecture or the need to perform real-time adversarial optimization in a challenge-response regime), *identity-based samples* could potentially scale quite easily (e.g., with pre-computed master embeddings used in a real time voice conversion system). Given the observed transferability of master voice samples obtained with voice cloning, it will be exciting to explore other speech representations and generators, such as disentangled representations [53] and voice conversion systems [54].

If our attack will become a viable threat vector, further work will be needed to devise countermeasures (e.g., by deploying an additional background population that reveals mass impersonation capabilities of the given sample). Overall, we believe our work will ultimately lead to a better understanding of the speech modality and more secure human-computer interaction.

APPENDIX A

To better understand the context of our study, this appendix collects a range of supplementary results and material:

- Detailed benchmark performance of the considered speaker encoders at different security thresholds.
- Receiver operation characteristics (ROC) curve for the speaker encoders considered in our study.
- Menagerie analysis conducted under each of the considered speaker encoders.
- Seed utterances ranking based on their impersonation rate against male and female users.
- The susceptibility of a utterance to achieve high impersonation rates between two speaker encoders.
- Impersonation rates of seed and master voices in case of multiple presentation attempts for NES-based attacks.
- Source code and data accompanying this paper available at https://github.com/mirkomarras/dl-master-voices.

REFERENCES

[1] A. K. Jain, D. Deb, and J. J. Engelsma, "Biometrics: Trust, but verify," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 4, no. 3, pp. 303–323, Jul. 2022.

[2] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, Mar. 2021.

[3] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, *Handbook of Fingerprint Recognition*, 2nd ed. Cham, Switzerland: Springer, 2009.

[4] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 74–99, Nov. 2015.

[5] T. Ammari, J. Kaye, J. Y. Tsai, and F. Bentley, "Music, search, and IoT: How people (really) use voice assistants," *ACM Trans. Comput. Hum. Interact.*, vol. 26, no. 3, pp. 17:1–17:28, 2019.

[6] M. B. Hoy, "Alexa, Siri, Cortana, and more: An introduction to voice assistants," *Med. Reference Services Quart.*, vol. 37, no. 1, pp. 81–88, 2018.

[7] Y. Chen et al., "Devil's whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices," in *Proc. USENIX Secur.*, 2020, pp. 2667–2684.

[8] Z. Wu et al., "SAS: A speaker verification spoofing database containing diverse attacks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4440–4444.

[9] X. Li, J. Zhong, X. Wu, J. Yu, X. Liu, and H. Meng, "Adversarial attacks on GMM I-vector based speaker verification systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6579–6583.

[10] F. Bimbot et al., "A tutorial on text-independent speaker verification," *EURASIP J. Adv. Signal Process.*, vol. 2004, no. 4, pp. 430–451, 2004.

[11] L. You, W. Guo, L.-R. Dai, and J. Du, "Deep neural network embeddings with gating mechanisms for text-independent speaker verification," in *Proc. Interspeech*, Sep. 2019, pp. 1168–1172.

[12] J. Yamagishi, T. Kinnunen, N. W. D. Evans, P. L. D. Leon, and I. Trancoso, "Introduction to the issue on spoofing and countermeasures for automatic speaker verification," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 4, pp. 585–587, 2017.

[13] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proc. ICASSP*, Mar. 2016, pp. 5115–5119.

[14] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4052–4056.

[15] Z. He, M. N. Rajput, and M. Ahamad, "Compromised computers meet voice assistants: Stealthily exfiltrating data as voice over telephony," in *Proc. 51st Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. (DSN)*, Jun. 2021, pp. 519–530.

[16] A. Roy, N. Memon, and A. Ross, "MasterPrint: Exploring the vulnerability of partial fingerprint-based authentication systems," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 9, pp. 2013–2025, Sep. 2017.

[17] P. Bontrager, A. Roy, J. Togelius, N. Memon, and A. Ross, "DeepMasterPrints: Generating MasterPrints for dictionary attacks via latent variable evolution," in *Proc. IEEE 9th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Oct. 2018, pp. 1–9.

[18] H. H. Nguyen, J. Yamagishi, I. Echizen, and S. Marcel, "Generating master faces for use in performing wolf attacks on face recognition systems," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2020, pp. 1–10.

[19] N. Yager and T. Dunstone, "The biometric menagerie," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 220–230, Feb. 2010.

[20] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst., Annu. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2014, pp. 2672–2680.

[21] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. 2nd Int. Conf. Learn. Represent., (ICLR)*, Banff, AB, Canada, Apr. 2014, pp. 1–14.

[22] M. Marras, P. Korus, N. Memon, and G. Fenu, "Adversarial optimization for dictionary attacks on speaker verification," in *Proc. Interspeech*, Sep. 2019, pp. 2913–2917.

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

[24] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, nos. 1–3, pp. 19–41, 2000.

[25] N. Dehak, P. J. Kenny, R. Dehak, D. Pierre, and O. Pierre, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[26] P. Rajan, A. Afanasyev, V. Hautamäki, and T. Kinnunen, "From single to multiple enrollment I-vectors: Practical PLDA scoring variants for speaker verification," *Digit. Signal Process.*, vol. 31, pp. 93–101, Aug. 2014.

[27] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with flexibility in utterance duration," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2017, pp. 584–590.

[28] T. Kaseva, H. K. Kathania, A. Rouhe, and M. Kurimo, "Speaker verification experiments for adults and children using shared embedding spaces," in *Proc. NoDaLiDa*, 2021, pp. 86–93.

[29] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 1021–1028.

[30] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Comput. Speech Lang.*, vol. 60, Mar. 2020, Art. no. 101027.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[32] Z. Wang, K. Yao, X. Li, and S. Fang, "Multi-resolution multi-head attention in deep speaker embedding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6464–6468.

[33] S. Yadav and A. Rai, "Frequency and temporal convolutional attention for text-independent speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6794–6798.

[34] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4879–4883.

[35] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification," in *Proc. Interspeech*, Sep. 2018, pp. 3573–3577.

[36] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. Interspeech*, Aug. 2017, pp. 999–1003.

[37] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using X-vectors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5796–5800.

[38] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5329–5333.

[39] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5791–5795.

[40] N. Carlini et al., "Hidden voice commands," in *Proc. USENIX Secur.*, 2016, pp. 513–530.

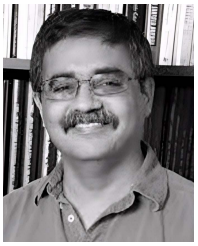
[41] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, May 2018, pp. 1–7.
[42] Y. Qin, N. Carlini, G. W. Cottrell, I. J. Goodfellow, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *Proc. ICML*, vol. 97, 2019, pp. 5231–5240.
[43] P. Neekhara, S. Hussain, P. Pandey, S. Dubnov, J. McAuley, and F. Koushanfar, "Universal adversarial perturbations for speech recognition systems," in *Proc. Interspeech*, Sep. 2019, pp. 481–485.
[44] Q. Wang, P. Guo, and L. Xie, "Inaudible adversarial perturbations for targeted attack in speaker recognition," in *Proc. Interspeech*, Oct. 2020, pp. 4228–4232.
[45] Y. Xie, Z. Li, C. Shi, J. Liu, Y. Chen, and B. Yuan, "Real-time, robust and adaptive universal adversarial attacks against speaker recognition systems," *J. Signal Process. Syst.*, vol. 93, no. 10, pp. 1187–1200, Oct. 2021.
[46] G. Chen et al., "Who is real bob? Adversarial attacks on speaker recognition systems," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2021, pp. 694–711.
[47] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Commun.*, vol. 66, pp. 130–153, Feb. 2015.
[48] A. Adler and S. Schuckers, *Security and Liveness, Overview*. Boston, MA, USA: Springer, 2009, pp. 1146–1152.
[49] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1983, pp. 804–807.
[50] Y. Jia et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proc. NeurIPS*, 2018, pp. 4485–4495.
[51] C. Jemine. *Real-Time Voice Cloning*. Accessed: Feb. 1, 2022. [Online]. Available: https://github.com/CorentinJ/Real-Time-Voice-Cloning, open source voice cloning system
[52] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6199–6203.
[53] J. Williams, Y. Zhao, E. Cooper, and J. Yamagishi, "Learning disentangled phone and speaker representations in a semi-supervised VQ-VAE paradigm," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 7053–7057.
[54] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *Proc. ICML*, vol. 97, 2019, pp. 5210–5219.
[55] D. Wierstra, T. Schaul, T. Glasmachers, Y. Sun, J. Peters, and J. Schmidhuber, "Natural evolution strategies," *J. Mach. Learn. Res.*, vol. 15, pp. 949–980, Mar. 2014.
[56] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.
[57] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, May 2001, pp. 749–752.
[58] Y. Jia et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," 2018, *arXiv:1806.04558*.
[59] Y. Wang et al., "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, Aug. 2017, pp. 4006–4010.
[60] N. Kalchbrenner et al., "Efficient neural audio synthesis," in *Proc. ICML*, vol. 80, 2018, pp. 2415–2424.
[61] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.
[62] G. Fenu, M. Marras, G. Medda, and G. Meloni, "Fair voice biometrics: Impact of demographic imbalance on group fairness in speaker recognition," in *Proc. Interspeech*, Aug. 2021, pp. 1892–1896.
[63] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," 2020, *arXiv:2005.07143*.
[64] M. Sagaya, S. V. Katta, and S. Umesh, "S-vectors: Speaker embeddings based on transformer's encoder for text-independent speaker verification," 2020, *arXiv:2008.04659*.

Mirko Marras (Member, IEEE) received the M.Sc. (summa cum laude) and Ph.D. degrees in computer science from the University of Cagliari, Italy, in 2016 and 2020, respectively. From 2020 to 2021, he spent 12 months as a Post-Doctoral Researcher at the Machine Learning for Education Laboratory, EPFL, Switzerland. He is currently an Assistant Professor with the Department of Mathematics and Computer Science, University of Cagliari. His research interests include data mining and machine learning techniques for user profiling and personalization, applied to the fields of behavioral analysis, education, business, entertainment, and social computing, with a focus on user impact. He is a member of (inter)national associations, including IEEE.

Paweł Korus (Member, IEEE) received the M.Sc. and Ph.D. degrees (Hons.) in telecommunications engineering from the AGH University of Science and Technology, Kraków, Poland.

He was a Post-Doctoral Researcher at the College of Information Engineering, Shenzhen University, China. He held appointments as an Assistant Professor with the Department of Telecommunications, AGH University of Science and Technology, and a Research Assistant Professor with the Center for Cyber-Security, New York University, Brooklyn, NY, USA. He is currently an Applied Scientist with Amazon. His research interests include multimedia signal processing, low-level vision, and security. In 2015, he received a Scholarship for Outstanding Young Scientists from the Polish Ministry of Science and Higher Education.

Anubhav Jain received the B.Tech. degree in electronics and communications engineering from the Indraprastha Institute of Information Technology, New Delhi, India. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Tandon School of Engineering, New York University, Brooklyn, NY, USA. He was a Research Intern at the Idiap Research Institute, Switzerland. His research interests include digital forensics, biometrics, and security.



Nasir Memon (Fellow, IEEE) received the B.Eng. degree in chemical engineering and the M.Sc. degree in mathematics from the Birla Institute of Technology and Science (BITS), Pilani, India, and the Ph.D. degree in computer science from the University of Nebraska.

He is currently a Professor with the Department of Computer Science and Engineering, Tandon School of Engineering, New York University, Brooklyn, NY, USA, and one of the Co-Founders of the Center for Cyber-Security. He has published over 250 papers in journals and conference proceedings and holds a dozen patents in image compression and security. His research interests include digital forensics, biometrics, data compression, network security, and human behavior. He is a fellow of SPIE. He has won several awards, including the Jacobs Excellence in Education Award and several best paper awards. He has been on the editorial boards of several journals. He was the Editor-in-Chief of IEEE TRANSACTIONS ON INFORMATION SECURITY AND FORENSICS.