AntiFake: Using Adversarial Audio to Prevent Unauthorized Speech Synthesis

Zhiyuan Yu
Washington University in St. Louis
St. Louis, USA
yu.zhiyuan@wustl.edu

Shixuan Zhai
Washington University in St. Louis
St. Louis, USA
zhais@wustl.edu

Ning Zhang
Washington University in St. Louis
St. Louis, USA
zhang.ning@wustl.edu

ABSTRACT

The rapid development of deep neural networks and generative AI has catalyzed growth in realistic speech synthesis. While this technology has great potential to improve lives, it also leads to the emergence of "DeepFake" where synthesized speech can be misused to deceive humans and machines for nefarious purposes. In response to this evolving threat, there has been a significant amount of interest in mitigating this threat by DeepFake detection.

Complementary to the existing work, we propose to take the preventative approach and introduce AntiFake, a defense mechanism that relies on adversarial examples to prevent unauthorized speech synthesis. To ensure the transferability to attackers' unknown synthesis models, an ensemble learning approach is adopted to improve the generalizability of the optimization process. To validate the efficacy of the proposed system, we evaluated AntiFake against five state-of-the-art synthesizers using real-world DeepFake speech samples. The experiments indicated that AntiFake achieved over 95% protection rate even to unknown black-box models. We have also conducted usability tests involving 24 human participants to ensure the solution is accessible to diverse populations.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Security and privacy** → **Human and societal aspects of security and privacy**.

KEYWORDS

Adversarial Machine Learning; Generative AI; Speech Synthesis; DeepFake Defense;

ACM Reference Format:
Zhiyuan Yu, Shixuan Zhai, and Ning Zhang. 2023. AntiFake: Using Adversarial Audio to Prevent Unauthorized Speech Synthesis. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23), November 26–30, 2023, Copenhagen, Denmark.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3576915.3623209

1 INTRODUCTION

Speech synthesis, commonly known as Text-to-Speech (TTS), refers to the generation of artificial human speech from textual input. Over the years, this technology has played a pivotal role across a wide spectrum of applications, ranging from accessibility aids for individuals with speech or hearing impairments to voice assistants in smart devices. The recent advent of Deep Neural Networks (DNNs) has further propelled growth in this field, leading to the emergence of highly realistic synthesized speech known as "DeepFake" audio. However, while these powerful systems have revolutionized human-computer interaction and are designed to improve lives, they also pose significant security risks due to their potential for nefarious applications.

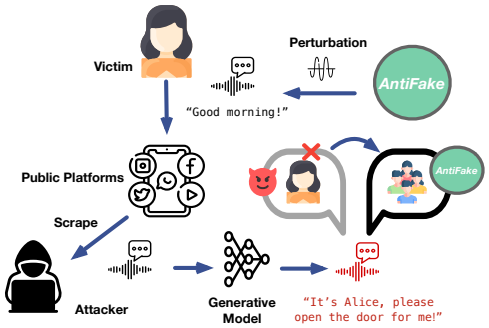

Figure 1: Overview of AntiFake.

Real-world Threats of DeepFake. The use of DeepFake speech in malicious attacks is not a distant possibility. It has been reported that fraudsters used DeepFake techniques to impersonate a CEO's voice and successfully swindled more than $243,000 through a phone call [47]. More recently in 2023, DeepFake audio has been utilized to breach bank accounts by convincingly impersonating account holders [15], as well as generate misinformation and hate speech in the guise of influential celebrities' voices [56, 65], resulting in widespread negative societal impacts. These incidents showed that contemporary speech synthesis techniques are already capable of deceiving both digital authentication systems and human auditory perception, highlighting the pressing need for effective countermeasures.

Existing Defenses and Inherent Limitations. In response to these newly emerged threats, existing research efforts have mostly been dedicated to developing detection methods for defense. More specifically, these approaches primarily focus on liveness detection [44, 46, 66] and acoustic signal analysis [3, 5], building upon insights from physical properties of human vocal systems (e.g., vocal cord vibrations [44] and articulatory gesture [66]), as well as lower-dimensional signal features (e.g., MFCC and signal power linearity degree [3]). Although these methods have demonstrated remarkable success in uncovering synthetic audio as a post-attack mitigation tool, AntiFake takes a departure from the existing line

CCS '23, November 26–30, 2023, Copenhagen, Denmark

ACM ISBN 979-8-4007-0050-7/23/11.
https://doi.org/10.1145/3576915.3623209

of research and focuses on preventing the attack to build multiple layers of defenses.

AntiFake - A Complementary Defense Layer via Preventing Unauthorized Synthesis. Motivated by this gap and the observation that the attacker has to rely on users' voice samples to generate DeepFake speech, we propose AntiFake to prevent unauthorized speech synthesis using stealthy adversarial perturbations. AntiFake takes a proactive approach, with the key idea of disrupting the synthesis process by deviating speaker embeddings used for speaker identity control in conditioned speech synthesis. Using AntiFake, users or platforms can protect speech samples before sharing them with the public. While the processed sample still sounds like the victim to humans, when it is used for speech synthesis by the attacker, the resulting synthetic speech would resemble others' voices rather than the victim's. Consequently, the generated DeepFake audio is less likely to deceive humans or machines for nefarious purposes.

Technical Challenges of AntiFake. Designing AntiFake presents four main technical challenges. First, we adhere to a practical setting where the user with AntiFake is unaware of the exact model employed by the attacker. Therefore, even black-box queries to optimize effective perturbations become infeasible, requiring adversarial examples to be transferrable to unknown models. To address this issue, we build on an assumption that adversarial synthesis models often have to share similarities with other robust encoders for improved efficacy, and adopt ensemble learning to cause significant deviations in those features. Intuitively, this approach ensures that perturbed embedding is always closer to embeddings from other speakers, compared to the original speaker. This principle forms the foundation for the two optimization mechanisms, which guide the exploration towards maximum embedding deviations.

Second, our investigation of state-of-the-art speech synthesizers reveals that they often employ multiple audio segments to facilitate robust embedding extraction. This characteristic renders traditional methods of directly calculating final embeddings less effective, which can trap the optimization in local optima. To overcome this problem, we employ a segment-based optimization strategy and introduce a weighted segment loss adaptation mechanism. This approach dynamically adjusts segment weights according to their contributions, enabling more refined perturbation generation by optimizing individual segments.

Third, it is crucial to maintain the perturbed audio samples' quality to ensure usability and prevent arousing an attacker's suspicion. Traditional L_p-based perturbation measurements are inadequate due to the gap between L_p distance and human auditory systems. Therefore, we turn to human perception principles, specifically frequency band sensitivity and masking effects. Building on psychoacoustic characteristics, we devise frequency penalties based on inverse sound pressure levels and incorporate computationally efficient SNR metrics to enhance imperceptibility.

Lastly, the primary goal of AntiFake to protect human perception necessitates human involvement in judging speaker identity deviation, as no computational model perfectly models auditory perception. However, using humans as iterative loss feedback is impractical and can significantly undermine usability. To this end, we design a human-in-the-loop workflow that minimally involves humans in non-technical tasks (i.e., only rating the difference in speaker identity) while maximizing the benefits of optimization. To further balance the computational embedding deviation and human judgment, we adapt the Analytic Hierarchy Process [53] to comprehensively balance the two and make informative decisions.

Experiments and Findings. To enable a comprehensive evaluation of AntiFake, we evaluated against five state-of-the-art synthesizers including one commercial product (i.e., ElevenLabs [18]), and three speaker verification systems including one commercial platform (i.e., Microsoft Azure). Furthermore, to facilitate a more realistic evaluation, we sourced real-world DeepFake sentences with categorization based on malicious intents. Our evaluation starts from a large-scale synthesis across diverse synthesizers, speakers, utterances, and speech content. Out of 60,000 synthesized speech samples, we filtered those that can bypass speaker verification systems and are perceptually similar to the victims' voices to form a set of high-fidelity DeepFake speech datasets. Built upon these, the evaluation of AntiFake achieved > 95% protection rate, and the optimized adversarial examples were shown transferrable to black-box commercial models. To validate usability, we further conducted usability tests based on the system usability scale (SUS) questionnaire, involving 24 human participants with diverse backgrounds.

Contributions. Our proposed AntiFake not only enhances security against DeepFake threats but also contributes to the ongoing battle against the spread of misinformation and impersonation. Our contributions are outlined as follows.

- We propose AntiFake, a proactive defense approach leveraging adversarial examples to disrupt unauthorized speech synthesis, such that the synthesized DeepFake audio does not resemble the victim's voice to both humans and machines.
- We develop a human-in-the-loop workflow of AntiFake to enable users to customize voices with minimal human effort. To address the challenge of unknown attackers' synthesizers, we adopt an ensemble learning approach on a set of combined state-of-the-art encoder models.
- We evaluate AntiFake against five contemporary synthesizers including one commercial product, and three speaker verification systems including one commercial platform. AntiFake achieves over 95% protection rate even against unseen commercial synthesizers. We also conducted usability tests involving 24 human participants.

2 BACKGROUND

Speech synthesis refers to the generation of human-like speech audio through computational systems. Traditional speech synthesis techniques, with origins tracing back to the 18th century, mainly relied on rule-based methods such as formant synthesis and diphone concatenation [36]. The recent advent of DNNs has revolutionized this field, also known as "DeepFake", with significantly improved quality and the ability to synthesize speech in zero-shot settings (i.e., the target speaker is not involved in the training of the generative models). As such, it further amplifies real-world threats where the attacker can generate audio using merely a few seconds of audio samples readily accessible on the Internet. In this study, we focus on such contemporary DNN-based models, with their structure summarized in Figure 2 and detailed in the following.

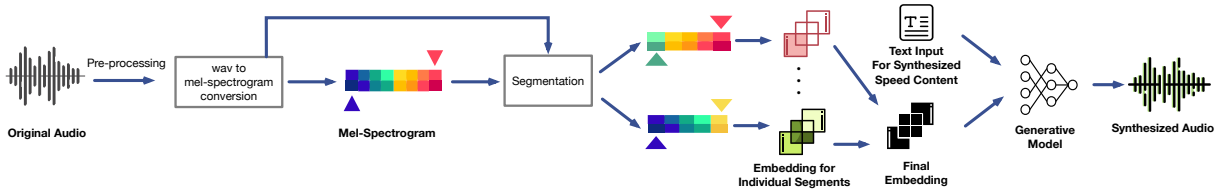

Figure 2: The general workflow of DNN-based speech synthesis process.

Preprocessing and Spectrogram Conversion. To ensure the optimal quality of synthesized audio and effective extraction of speaker characteristics, typical audio pre-processing methods include resampling to obtain a consistent sample rate, normalizing volume to establish a uniform loudness level, and trimming silence in speech. The preprocessed audio waveform is subsequently transformed into mel-spectrograms, a time-frequency representation that effectively captures the essential features of the speech signal.

Speaker Embedding Encoding. The key component that characterizes the efficacy of speech synthesis is the speaker embeddings extracted via encoders. These embeddings capture the features related to speaker identity and generally possess a fixed length. For effective synthesis, a major goal of the constructed embedding space is to ensure that embeddings for the same speaker exhibit high similarity irrespective of the content while maintaining a considerable distance from embeddings of other speakers. Therefore, contemporary systems often partition audio and associated mel-spectrograms into smaller segments, thereby more accurately capturing the diverse phonetic and prosodic attributes of the speaker's voice. As such, this approach aims to address potential inaccuracies that may arise from averaging various features across long time frames.

Conditional Synthesis with Textual Inputs. Similar to speaker encoding, the speech content controlled by textual inputs is transformed into linguistic features or phoneme-based representations. Such encoded content together with speaker embedding will go through DNN models to synthesize raw audio. After subsequent postprocessing, this synthesis process conditioned on speaker embedding will ultimately produce audio that closely mimics the target speaker's vocal characteristics with high fidelity.

3 RELATED WORK

3.1 Adversarial Audio Examples

Adversarial examples have emerged as a significant threat to machine learning systems, with the seminal work of Goodfellow et al. [20] revealing the vulnerability of image classification models to subtle perturbations that lead to misclassification. Beyond images, adversarial machine learning has since evolved and expanded to audio-based machine learning systems, such as automatic speech recognition and speaker identification [62]. Carlini et al. [7] introduced the hidden voice command attack where the obfuscated audio piece can be deciphered as commands by machines while remaining unintelligible to humans. Yuan et al. [64] investigated DNN-based speech recognition models, and devised perturbations embedded within musical compositions to convey malicious commands. Moreover, Carlini et al. [8] focused on clean speech, and proposed a white-box attack using gradient descent on the CTC loss. Subsequent advancements in the field include Metamorph [11], which achieved over-the-air delivery of adversarial audio by incorporating channel impulse responses and frequency responses. Schönherr et al. [42] improved attack stealthiness by incorporating a psychoacoustic model, while Abdullah et al. [1] improved attack generalizability by exploiting the model-agnostic audio preprocessing stage. On the other hand, [10, 33, 34] explored adversarial attacks targeting speaker identification, with the optimization objective of maximizing the confidence score of the target speaker. Recently, Yu et al. [63] proposed semantic audio attacks against both speech transcription and speaker recognition systems, where they departed from L_p-based perturbations and instead manipulated prosody features to maximally retain speech quality and naturalness. While existing work primarily focuses on speech and speaker recognition systems, the target of AntiFake is the speech synthesis generative model that possesses completely different architectures, which therefore requires new attack designs.

3.2 Adversarial Examples for Defense

Recent advances in machine learning also inadvertently empower attackers to exploit DNN models for malicious intents, which inspired defenses that leverage adversarial examples for mitigation. Within this context, Fawkes [43] was proposed as a poisoning-based defense against unauthorized facial recognition that threatens people's privacy. By altering pixel values within facial regions, face recognition models trained on such data will misidentify protected users during inference. Following this work, Cherepanova et al. [12] proposed an improved approach by formulating more practical settings for face recognition models and utilizing a larger dataset. In the audio domain, Abdullah et al. [2] proposed to disrupt speech and speaker recognition by decomposing signals and filtering out non-essential components for human comprehension. Another related work [26] conducted initial explorations of compromising voice conversion models, however, it faces several limitations within our context. First, their attack was designed to target specific models, resulting in model-specific adversarial examples. However, users cannot predict the exact models that attackers may employ, which therefore necessitates a new mechanism generalizable to a variety of synthesis models to best protect users. Second, contemporary speech synthesis models have evolved with more advanced architectures and techniques (e.g., attention mechanism), which demand new attack strategies to ensure effectiveness. Third, [26] requires tunning parameters of the attack algorithm, which poses challenges for users from diverse backgrounds and restricts customizability. In our work, we introduced a human-in-the-loop mechanism to automatically adjust the optimization process, striking a balance

between audio quality and protection strength based on user customizations. More details of our designs are discussed in Section 5.

3.3 Defense against DeepFake Speech

The increasing threats brought by speech synthesis have inspired several existing defenses, primarily focusing on DeepFake speech detection. There are two main approaches: liveness detection based on physical properties and signal analysis based on synthesis artifacts. More specifically, liveness detection methods leverage the unique acoustic characteristics induced by the physical aspects of human speech such as pop noise caused by human breadth [46], articulatory gesture [66], or vibration of human vocal cords [44]. In contrast, signal analysis methods [3, 5] are designed to examine acoustic signals and extract lower-dimensional features for classification, where the synthesis artifacts are generally captured and therefore distinguish them from natural speech. Such features can either be pure signal properties (such as MFCC and signal power linearity degree) [3] or those associated with physical aspects such as vocal tract structure [5]. While they exhibit remarkable performance, all of these works aim at DeepFake audio detection deployed after receiving suspicious speech audio on the user side. In contrast, AntiFake targets the speech synthesis stage on the attacker side. As such, AntiFake aims to prevent attackers from synthesizing convincing audio in the first place, offering a complementary approach to existing defenses and providing an additional layer of protection.

4 THREAT MODEL

4.1 Attacker Motivation and Assumptions

Attack Motivation. The attacker aims to synthesize DeepFake speech that convincingly mimics the target speaker for a variety of malicious purposes. We outline four motivations most commonly seen in real-world scenarios, with the first three focused on deceiving humans and the last one on tricking machines.

(1) Conduct Financial Scam. Attackers may create DeepFake speech to impersonate company officials, bank executives, or individuals with a close relationship to the victim to conduct financial scams. As a real-world motivating example, in 2019, the attacker used DeepFake speech to impersonate the boss of a CEO, which led to the scamming of $243,000 transferred to the attacker's account [47].

(2) Compromise Safety and Privacy. By impersonating a trusted individual such as a colleague or supervisor, attackers may trick victims into taking unsafe measures or sending sensitive data (e.g., access tokens, passwords, personal identifiable information (PII)). Such examples existed in the real world during the Covid-19 pandemic, where DeepFake speech has been used by attackers to apply for remote positions that can access enterprise secrets such as PII, financial data, and corporate IT databases [14].

(3) Spread Hate Speech or Misinformation. Distributing hate speech or misinformation through DeepFake speech can have far-reaching societal impacts. By attributing false information to influential individuals or organizations, attackers can manipulate public opinion, sow discord, and exacerbate tensions. Recently in 2023, it was reported that attackers cloned the voices of celebrities to spread negative content such as erotica, hate speech, and misinformation [56].

(4) Bypass Voice-based Authentication. Voice-based authentication has been increasingly deployed in safety-critical systems, relying on the unique characteristics of an individual's voice to verify their identity and grant access. However, the advanced DeepFake techniques that are optimized for imitating those features can bypass such authentication schemes. Such threats have been demonstrated in a real-world case that happened in 2023, where an attacker managed to break into a bank account using AI-synthesized voices [15].

Built upon the above intentions, we compiled a list of example sentences sourced from the real world for synthesis in our experiments, listed in Table 6 in the appendix.

Attacker Assumption. We assume the attacker is a third-party entity without direct access (i.e., record speech physically) to the target user's speech samples. Instead, they will collect data from public-domain resources, including but not limited to social media, websites, and video streaming platforms. The attacker has moderate computational resources that enable them to run existing speech synthesizers, or alternatively train a new model (either from scratch or fine-tune models) using publicly available speech data. Following recent studies in DeepFake audio [5, 60], we assume the attacker employs state-of-the-art synthesizers and uses the victim's speech samples to conduct zero-shot speech synthesis. Additionally, we assume that adversarial models often have to share similarities with other robust encoders for improved efficacy, which is reflected as transferability and validated in ablation studies (Section 6.7).

4.2 System Goals and User Assumptions

User Objective and Assumptions. The ultimate objective of the user is to share their speech audio online without inadvertently helping attackers to synthesize DeepFake speech for malicious purposes. To do so, users employ AntiFake to process audio prior to publishing. AntiFake leverages adversarial optimization to add subtle perturbations to the original audio to be shared, such that the synthesized audio will not resemble their voice to both humans and machines (i.e., voice-based authentication systems). In this study, we assume that neither the user nor AntiFake has knowledge of the specific synthesis model employed by attackers.

Primary System Goals of AntiFake. To best provide protection to users with minimal changes to the original media, the key system goals of AntiFake are two-fold. First, The speech synthesized using the adversarial speech audio should produce speech audibly different from the user. Second, The perturbations should be imperceptible to human perception, and the perturbed speech audio should still sound natural and high-fidelity. The core technical designs of AntiFake closely align with these objectives, which are detailed in Section 5.

5 ANTIFAKE DESIGN

5.1 Overview and Technical Challenges

AntiFake is designed to align with the two system goals. The major goal is to change the speaker identity of synthesized speech. As discussed in Section 2, the fundamental objective of robust and effective speech synthesis is to generate a convincing voice that remains consistent regardless of the speech content. Due to this reason, speech embeddings dominantly determine the voice identity

by nature. Inspired by this principle, the key intuition of AntiFake is to disrupt embedding space with perturbations added to the original audio samples. It might appear straightforward to simply add high-magnitude noises throughout the entire audio sample, which could lead to a totally different speaker embedding. In this case, the user is "perfectly" protected but the original audio is completely ruined. As such, achieving the dual objectives of strong protection and quality preservation requires a delicate balance in the optimization of perturbations.

To achieve this, we proposed two optimization mechanisms integrated in AntiFake. Specifically, the *threshold-based* approach aims to deviate embedding away from the original speaker based on thresholds, while the *target-based* method aims to shift embedding close to a target speaker who is not the original speaker. These two approaches are individually advantageous in different aspects, which are further discussed in Section 5.4. But in both cases, the synthesized audio will ideally sound like someone else with respect to both humans and machines, thereby mitigating the threats brought by DeepFake audio.

We summarize the technical challenges into four main aspects.

C1 - Synthesizer Employed by Attackers is Unknown. We make a practical assumption that users are unaware of the exact synthesizer utilized by the attacker, which is the key challenge that requires generalizable perturbations that can transfer to an unknown model. To solve this problem, we build on existing findings [30] that robust encoders exhibit similarities in the extracted characteristics and latent space boundaries (i.e., consistently projecting different speakers into distinct embedding areas while mapping the same speaker to identical regions). As such, we reasonably hypothesize that the embeddings from unknown encoders are different when the corresponding speech samples are sufficiently different in inherent acoustic properties. Therefore, we adopted an ensemble learning approach, incorporating state-of-the-art encoders to optimize perturbations generalizable across various synthesizers.

C2 - Handling Robust Embeddings Derived from Segments. For robust speaker embedding extraction, state-of-the-art speech synthesizers typically first generate partial embeddings from speech segments and subsequently integrate them into a final embedding (Section 2). To this end, our initial exploration revealed that directly deviating the final embedding is susceptible to local optima, as the impact of a partial embedding not close to the target is mitigated by other partials that may have attained better optimal points. To address this problem and improve effectiveness, we propose a weighted segment-based raw embedding deviation technique incorporating weighted loss calculation. This approach allows for more fine-grained control and ensures that each partial embedding contributes effectively to the overall perturbation.

C3 - Requirement on Stealthy Perturbations. Making perturbations imperceptible is also a key requirement to ensure usability and effectiveness. However, we found that L_p-norms commonly employed by existing literature are less effective, and the resulting perturbations are still audible. To overcome this limitation, we exploit the human auditory system's diverse sensitivity to audio frequencies, and penalize perturbations with nuanced per-frequency-band gains inversely proportional to their sound pressure levels.

C4 - Difficulty of Obtaining Continuous Feedback from Humans. A primary objective of AntiFake is to protect human perception against DeepFake speech. While there is no perfect model that can replace the human auditory system, we design a human-in-the-loop approach. However, obtaining continuous feedback from humans during optimization is labor-intensive and can significantly undermine usability. As such, how to minimize human efforts while retaining effectiveness is a key challenge in designing the system. To this end, we designed a process where the user is only asked to subjectively judge the speaker identity dissimilarity of the speech samples. In order to balance human judgment and computed embedding deviation, we adopt Analytic Hierarchy Process (AHP) [53] to comprehensively incorporate the two. The workflow involving users in the loop is described in the following.

5.2 AntiFake with Human-in-the-loop

The protection objectives of AntiFake necessitate the involvement of human efforts to validate deviations in speaker identity within the synthesized audio. For usability, the designed process adheres to two principles: first, tasks assigned to users must be accessible to those with minimal prior knowledge, and second, human efforts should be minimized while maximizing the benefits of optimization. The overall workflow of AntiFake can be summarized as follows:

Stage #1 - Speech Upload and Embedding Extraction. The initial step starts with users uploading the speech audio they wish to protect. This audio is then pre-processed, and speaker embeddings are extracted utilizing an ensemble of encoders. More details of ensembled encoders are described in Section 5.5. Both the speech audio and associated embedding are used for further analysis.

Stage #2 - Target Selection with Analytic Hierarchy Process. In this step, AntiFake searches its database consisting of public speech corpora such as VCTK [61] to identify potential targets. It randomly selects five sentences each from a unique speaker. These sentences are used to compute speaker embeddings, which are ranked based on their distance from the original embedding of the user's uploaded speech. To further ensure that target speakers are perceptibly different from the user, the user is prompted to listen to these audio samples and rate the perceived identity dissimilarity on a scale of 1 to 5, with 5 representing the highest dissimilarity and 1 denoting the least. Due to the gap between embedding representation and human auditory perception, conflicts may arise between embedding-based rankings and human scores. To this end, we model it as a multi-criteria decision-making problem and adopt the Analytic Hierarchy Process to determine the appropriate target. In our context, we model computational dissimilarity and human scores as the decision-making criteria, and the five sentences are the decision alternatives. In this way, our approach considers relationships among candidate sentences through pairwise matrix calculations and integrates human perception and embedding deviations with the aggregated priority vectors. In the end, the target sentence (and target speaker) is selected with the largest deviation in both human perception and computational representations.

Stage #3 - Optimization and User Perceptual Validation. The obtained speech audio and target sentence (which also derives the threshold by calculating embedding distance) are subsequently used for optimization as detailed in Section 5.4. The finalized perturbed

speech audio is then used to synthesize speech and sent to the user for listening tests as done in Step #2. If the resulting synthesized speech does not sound sufficiently different from the user, the user is provided with the opportunity to either proceed with optimization or restart with automatically adjusted hyperparameters.

5.3 Formulating the Optimization Problem

We first define and formulate the problem as follows:

- $\mathbf{x_U}$: The original speech audio of the user $\mathcal{S}_\mathcal{U}$ to be protected.
- $\mathbf{x_T}$: The target speech audio of speaker $\mathcal{S}_\mathcal{T}$.
- $\delta_{\mathbf{x_U}}$ The perturbations generated by AntiFake.
- $G(t, x)$: The speech synthesis generative model that takes speech x and textual content t.
- $g(x)$: The encoder model that extracts speaker embeddings from the input audio x.
- $H(x)$: The human perception function that determines the speaker identity with given speech x.
- $SV(x)$: The speaker verification system that takes speech x.
- $D(z, z^{'})$: The distance function measuring the difference between two speaker embeddings z and $z^{'}$.
- $M(\delta)$: The function measuring the perturbation magnitude with respect to human perception.

Given the original speech audio $\mathbf{x_U}$, AntiFake aims to deviate the speaker embedding away from the original one, formulated as:

$$\underset{\delta_{\mathbf{x_U}}}{\text{maximize}} \quad D\left(g(\mathbf{x_U}+\delta_{\mathbf{x_U}}), g(\mathbf{x_U})\right) - \alpha M(\delta_{\mathbf{x_U}}), \tag{1}$$

$$\text{subject to} \quad H(\mathbf{x_U}) \approx H(\mathbf{x_U}+\delta_{\mathbf{x_U}}) \tag{2}$$

$$H(G(t, \mathbf{x_U})) \neq H(G(t, \mathbf{x_U}+\delta_{\mathbf{x_U}})) \tag{3}$$

$$SV(G(t, \mathbf{x_U})) \neq SV(G(t, \mathbf{x_U}+\delta_{\mathbf{x_U}})) \tag{4}$$

where α is a positive hyperparameter that balances the imperceptibility of perturbations and the strength of the protection.

5.4 Two Schemes for Controlled Optimization

While the ultimate goal as indicated in Eq. 1 is to sufficiently deviate the speaker embedding, however, it is not always the bigger the deviation the better the effectiveness. For instance, in extreme cases, the deviated embedding grows to abnormally high values, and speech synthesis will completely fail (i.e., complete silence or pure whistling noises) which can easily draw attention from the attacker. As such, we consider it necessary to add reasonable control to the deviation magnitude (i.e., protection strength). To achieve this goal, we developed two mechanisms, one based on threshold and another based on existing embeddings belonging to other speakers.

Threshold Based. The protection strength can be directly controlled with a reasonable threshold, denoted as $\mathcal{T}$. In this context, we construct the loss function for embedding/identity shifting as:

$$\mathcal{L}_{\text{identity}}(\delta) = \text{ELU}[\mathcal{T} - D\left(g(\mathbf{x_U}+\delta\mathbf{x_U}), g(\mathbf{x_U})\right), a] \tag{5}$$

$$\text{where} \quad \text{ELU}(x, a) = \begin{cases} x, & \text{if } x > 0 \\ a(e^x - 1), & \text{if } x \leq 0 \end{cases}$$

The exponential linear unit (ELU) function is adopted to drive the optimization process into encouraging the embedding deviation towards the threshold. It offers two key advantages. First, it effectively shrinks the difference between the threshold and embedding deviation when the embedding difference is lower than $\mathcal{T}$. The protection strength reflected by $\mathcal{T}$ is only achieved when these two are equal. Second, it tolerates embedding deviations beyond $\mathcal{T}$, but only to a limited extent controlled by the alpha parameter (a in Eq. 5), which governs the negative saturation region of the function and determines the slope of the function in the negative region.

Target Based. In addition to a predefined threshold for controlling deviation magnitude (and the protection strength), an alternative approach is to guide the optimization towards a known speaker embedding with a completely different identity. A potential advantage of this method is to better preserve the naturalness of the deviated speaker embedding (since the target speaker exists) and the resulting synthesized speech. In this case, the loss function is:

$$\mathcal{L}_{\text{identity}}(\delta) = D\left(g(\mathbf{x_U}+\delta\mathbf{x_U}), g(\mathbf{x_T})\right), \tag{6}$$
$$s.t. \begin{cases} \mathbf{x_U} \in \mathcal{S}_\mathcal{U}, \mathbf{x_T} \in \mathcal{S}_\mathcal{T}, \\ \mathcal{S}_\mathcal{U} \neq \mathcal{S}_\mathcal{T}, \end{cases}$$

Note that the intention of introducing a target speaker is not to enforce the synthesis into producing speech sounds like $\mathcal{S}_\mathcal{T}$, which is unnecessary in our context. Instead, it aims to guide the optimization towards a more natural embedding and DeepFake speech.

The intuitive trade-off between the two schemes is that, the threshold-based method allows for a larger solution space, providing more flexibility in the optimization process. However, the resulting embedding is less controlled and may lead to artifacts in the synthesized speech. In contrast, the target-based method ensures more natural synthesis results by driving the optimization toward an existing speaker embedding. However, targeting a specific speaker actually introduces an additional constraint and significantly narrows the solution space. In cases where such an optimal solution is hard to achieve, it is possible that the optimization process may become stuck in local optima or require greater perturbations that degrade the original audio quality. Such trade-offs are further experimentally studied in Section 6.

5.5 Improving Efficacy and Imperceptibility

Improve Robustness via Ensembled Encoders. The vast majority of adversarial audio examples are designed to target specific text-to-speech or speaker recognition models [7, 8, 34, 63, 64]. Under the context of AntiFake as a defense, however, the synthesis or encoder model employed by the attacker is unknown, even unavailable to the users for black-box query. Such a knowledge gap makes it fundamentally challenging to develop perturbations that effectively disrupt the diverse embedding spaces of unknown models.

However, regardless of the model structures and parameters, all of them are designed to robustly extract unique acoustic features belonging to individual speakers. As such, we reasonably hypothesize that as long as the inherent acoustic features in the speech audio are sufficiently altered by the perturbations, the produced speaker embedding will deviate from the original one, even in the case of an unknown model. To ensure that the perturbations effectively alter the speaker-specific features, we surveyed state-of-the-art speech synthesis models and selected a total of 4 diverse encoders as summarized in Table 1. The chosen encoders exhibit a range of diverse properties, including different architectures from convolutional

Table 1: Properties of the ensembled encoders.

Encoder	Architecture	# Layer	Emb. Size	Dataset / # Speaker
AdaIN [13]	VAE	32	128	VCTK / 109
GE2E [58]	LSTM	3 (768 cells)	256	Custom / 18000
H/ASP [23]	ResNet	34	512	VoxCeleb 2 / 6000
ViT [17]	Transformer	31	1024	Proprietary / -

Embed. Size = Embedding Size

models (e.g., ResNet) to sequential models (e.g., LSTM), varying embedding sizes reflecting different levels of feature richness, and distinct training data to cover a more comprehensive set of individual speakers. In this way, we aim to derive comprehensive feature extraction to best benefit transferability across diverse synthesizers.

As the second step, we adopt an ensemble learning approach for altering these features and the embedding space in a robust manner. To achieve this, the loss values derived from individual encoders are jointly backward to optimize perturbations on the speech audio. With this approach, AntiFake ensures that the resulting adversarial examples are more likely to be effective against a wide range of unknown models employed by potential attackers.

Improve Effectiveness via Weighted Segment Loss Adaptation. The state-of-the-art speech synthesizers are designed to generate the ultimate speaker embedding by combining partial embeddings derived from multiple audio segments. This process ensures robust feature extraction and consequently better performance. Due to this reason, directly calculating the distance between the ultimate embeddings of the user and target as the optimization term may appear intuitive, however, we found that this approach can be less effective and prone to getting trapped in local optima. This limitation arises because partial embeddings in suboptimal situations are counterbalanced by other segments that may have already achieved optimal values. As a result, the entire embedding becomes difficult to adjust further. To address this challenge, we propose an inverse proportional loss adaptation scheme, which adaptively assigns weights to the loss terms of individual audio segments based on their contributions. The adaptive weight is designed as:

$$\mathcal{W}_i = \frac{[d\left(g(\mathbf{x_U}^i + \delta^i), g(\mathbf{x_U}^i)\right) + \epsilon]^{-1}}{\sum_{k=1}^{K} 1/[d\left(g(\mathbf{x_U}^k + \delta^k), g(\mathbf{x_U}^k)\right) + \epsilon]}, \tag{7}$$

where $\mathcal{W}_i$ is the weight corresponding to the i-th audio segment $\mathbf{x_U}^i$, and K is the total number of segments. The distance $d(\cdot)$ is calculated by the sum of the absolute values of the delta between all the individual elements within the two (partial) embeddings:

$$d(\mathbf{e_1}, \mathbf{e_2}) = \sum_{i=1}^{n} \left| \sum_{j=1}^{m} \left(\mathbf{e}_{i,j}^{(1)} - \mathbf{e}_{i,j}^{(2)} \right) \right|, \tag{8}$$

where $\mathbf{e_1}$ and $\mathbf{e_2}$ are the two embeddings being compared, while $n \times m$ represents the total dimensionality of the embedding. As such, the weighted distance for the target-based method is:

$$D(\delta) = \sum_{k=1}^{K} \mathcal{W}_k \cdot d\left(g(\mathbf{x_U}^k + \delta \mathbf{x_U}^k), g(\mathbf{x_T})\right), \tag{9}$$

and that of the threshold-based method follows a similar formula.

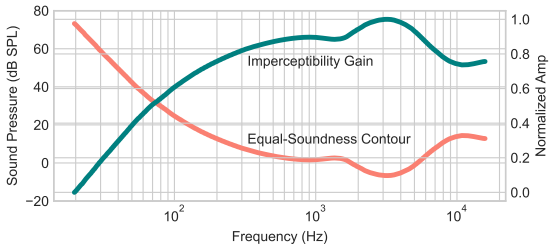

Figure 3: The equal-soundness contour and imperceptibility gain curve. The gain is calculated as the inverse linear normalization of the soundness pressure levels.

Improve Imperceptibility via Frequency Penalty. The imperceptibility of perturbations is a critical requirement, as we aim to minimally impact the normal use of the audio and avoid raising suspicion from the attacker. Existing work has primarily focused on minimizing the L_p-norm as the measurement of the perturbation magnitude [8, 11, 64]. While this approach indeed restricts the overall loudness, perturbations optimized in this manner are still audible due to the gap between L_p-norm and human perception.

To further enhance imperceptibility, we leverage the diverse sensitivity of the human auditory system to different audio frequencies [22]. This discrepancy stems from the resonance of the ear canal and the transfer function of the ossicles in the middle ear, making human hearing most responsive to certain frequencies [24]. Despite this, existing literature has explored more nuanced human hearing sensitivity related to frequencies. In our study, we build upon the equal-loudness contour [48], which is one of the most widely acknowledged assessments of sound pressure level (SPL) across the frequency spectrum. The raw data was collected through psychoacoustic experiments, which capture the gain as a linear reflection of human perception across different frequencies. We leverage this principle to linearly map sound pressure levels from -20 dB_{SPL} to 80 dB_{SPL} to imperceptibility gains ranging from 1 to 0 for different frequency bands. The original contour curve and the corresponding frequency gain are presented in Figure 3. In our context, a higher gain corresponds to a frequency band that is more challenging to perceive by humans, as reflected in higher sound pressure levels in the contour. Therefore, the perturbations in the frequency band, when multiplied by this gain, will be preserved to a greater extent because they are more "difficult" to perceive by humans. Formally, the frequency penalty term is defined as follows:

$$\mathcal{L}_{\text{percept}}(\delta) = \sum_{k=1}^{K} W(f_k) \cdot |\hat{\delta}(f_k)|, \tag{10}$$

where K is the number of frequency bands, f_k represents the center frequency of the k-th frequency band, $W(f_k)$ is the imperceptibility gain function for the frequency band f_k, and $\hat{\delta}(f_k)$ is the frequency strength of the perturbation in the k-th frequency band.

However, this measurement only considers the perturbation itself and disregards the original speech context. For instance, the same perturbations added to a loud speech sample will be less perceptible than those added to a gentle speech. To this end, we additionally incorporate the Signal-to-Noise Ratio (SNR) as the complementary metric. From the perspective of human auditory

perception, SNR is closely related to the *masking effect*, where the presence of a dominant signal can conceal the perception of weaker signals (i.e., perturbations) [6]. Moreover, it is a computationally efficient measure that facilitates faster optimization and thus improved usability. Finally, the $\mathcal{L}_{\text{identity}}(\delta)$, $\mathcal{L}_{\text{percept}}(\delta)$, and SNR are jointly optimized to effectively deviate the speaker embedding while maintaining changes to the original audio imperceptible.

6 EXPERIMENTS AND EVALUATION

6.1 Speech Synthesis Models

To best resemble a strong attacker with advanced DeepFake techniques, we focus on DNN-based speech synthesis models with state-of-the-art zero-shot capabilities. We manually examined the performance of the surveyed models, and five synthesizers were chosen. These include four open-source platforms, *SV2TTS* [58], *YourTTS* [9], *TorToiSe* [4], *Adaptive Voice Conversion (AdaptVC)* [39], and one commercial product named *ElevenLabs* [18].

SV2TTS. SV2TTS [58] is a three-step voice conversion system: LSTM-based speaker encoding [58], Tacotron 2 as synthesis network [45], and WaveNet vocoder [54]. We followed a well-established implementation on GitHub [29] and used the pre-trained models.

YourTTS. YourTTS [9] is a zero-shot multi-speaker TTS system with multilingual capabilities, introduced in 2022. It consists of three main components: the H/ASP model [23] as the speaker encoder, a custom VITS-style model as the decoder, and HiFi-GAN [32] as the vocoder. For implementation, we use the YourTTS model featured by Coqui, a library built on the latest TTS research.

TorToiSe. TorToiSe [4] is a recent open-source project on GitHub. Its overall architecture is inspired by OpenAI's DALLE model. The system is highly complex incorporating five separately-trained large models. For instance, it applies a GPT-2 model to predict token codes that represent highly-compressed audio data, a diffusion decoder to synthesize mel-spectrograms, and a UnivNet vocoder [28] to transform spectrograms into acoustic waveform.

Adaptive Voice Conversion. AdaptVC [39] possesses an autoencoder architecture with customed DNNs using instance normalization layers. The models are trained on CSTR VCTK Corpus [61].

ElevenLabs. ElevenLabs [18] is a newly-funded commercial product specializing in versatile AI speech technologies including TTS. Since the release of its beta platform in 2023, it has gained significant popularity and has even been misused in DeepFake to disseminate misinformation and hate speech [56]. The TTS product used in our study is a completely black-box system.

In addition, we also explored some other speech synthesis platforms such as Uberduck.ai [52]. However, they were excluded from the main study due to their relatively lower speech synthesis performance or limited capabilities in enabling customized speakers.

6.2 Speech Corpus Datasets

Due to ethical considerations, we create a custom dataset using speech audio samples from public datasets to represent the victim's and target's voices. The source datasets include VCTK [61], LibriSpeech [37], Speech Accent Archive (SAA) [59], and TIMIT [19].

Table 2: The attributes of the selected speech samples.

Source	# Speakers (M/F)	Age Range	Accent	Average Length
VCTK	25 (8/17)	18-38	American, Australian, Canadian, English, Indian, Irish, NewZealand, NorthernIrish, Scottish, SouthAfrican, Welsh.	4.78 s
Librispeech	25 (9/16)	-	-	4.70 s
SAA	25 (8/17)	19-66	Albanian, Cantonese, Czech, Dutch, English, French, Hebrew, Hindi, Italian, Japanese, Korean, Mandarin, Polish, Russian, Spanish, Tagalog, Taiwanese, Turkish, Ukrainian, Vietnamese, Xiang, Xasonga.	5.74 s
TIMIT	25 (12/13)	-	American (New England, Northern, North Midland, South Midland, Southern, New York City, Western, Moved around)	3.11 s

VCTK. The CSTR VCTK [61] dataset comprises over 44 hours of audio data produced by 109 English speakers with various accents. The recorded phrases are sourced from diverse text materials, including newspapers, linguistic texts, and phonetically-rich sentences.

LibriSpeech. LibriSpeech [37] is a large-scale corpus containing approximately 1,000 hours of English speech. We use the test-clean subset, which contains spoken phrases from 40 English speakers.

Speech Accent Archive. The SAA dataset [59] consists of a consistent set of spoken phrases in English, recorded by 2140 speakers from 177 different countries that represent 214 native languages.

TIMIT. TIMIT [19] includes audio recordings from 630 speakers of eight major dialects of American English, with each speaker reading ten phonetically rich sentences.

As summarized in Table 2, the speech samples were selected to cover a comprehensive set of speech content, sample lengths, as well as speakers with a wide range of genders, age groups, regions, and accents. As a result, our custom dataset comprises 100 speakers, with 25 from each source corpus, featuring five short (3-4s) to long (6-7s) audio samples per speaker.

6.3 Speech Content of DeepFake Audio

The speech content is equally important to cause negative impacts. As this factor has been overlooked in the past, we address this gap by compiling a list of sentences that can be exploited using DeepFake voices in various scenarios described in Section 4. To best simulate real-world DeepFake threats, we sourced a portion of sentences from Twitter instances and bank account verbal passwords [25, 57]. To obtain an appropriate set of sentences, we employed the Delphi method as a systematic consensus-building process. Initially, each author submitted their proposed list of sentences, with explanations regarding potential harmfulness, contextual relevance, and possible misuse scenarios. Subsequently, these sentences were consolidated and subjected to three rounds of iterative refinement. The finalized list of sentences can be found in Table 6 in the appendix, categorized based on the malicious intent that attackers aim to convey.

6.4 Speaker Verification Systems and Setup

We focus on three speaker verification systems; two open-source models (GMM-UBM and ivector-PLDA) and one commercial platform (Microsoft Azure) with configurations detailed as follows.

GMM-UBM and ivector-PLDA. Both GMM-UBM [40] and ivector-PLDA [16] are speaker recognition systems widely studied in existing research [10, 60]. We used their open-source implementation featured by Kaldi toolkit [38]. In our study, they were used to perform speaker verification (SV) tasks that verify whether a voice sample matches its labeled speaker. We first enrolled all target speakers from our custom dataset with their authentic speech samples, then evaluated any arbitrary utterance against its labeled speaker model for checking. Following existing work [60], we conducted batch testing and adjusted the output threshold of models using the Equal Error Rate (EER) criterion, defined as the error rate at which a system's false positive and false negative rates are equal. Specifically, we fine-tuned and selected the threshold while maintaining EER, by passing a controlled group of voice samples labeled with their real speakers in addition to our testing group. The success rate was subsequently calculated using the updated threshold.

Microsoft Azure. Adopted by the International Standards Organization (ISO), Microsoft Azure's cloud platform offers a commercialized solution for speaker recognition tasks. Similar to the open-source platforms, the verification process entails speakers enrolled with their authentic speech data, and the speaker verification queries can be made from its API.

6.5 Experimental Methodology

Large-scale Synthesis and Target Selection. To design comprehensive experiments, we leverage the synthesizers, speech samples, and DeepFake content discussed previously. Specifically, we employed an iterative process wherein each speaker was considered the source speaker (victim) and looped through each of their utterances. The synthesis was performed using each of the five synthesizers detailed in Section 6.1, resulting in a large-scale synthesis process that produced 60,000 synthesized speech samples ($100 \times 5 \times 24 \times 5$). From this extensive collection, we identified samples that successfully bypassed at least one speaker verification system (Section 6.4) while maintaining high-fidelity and perceptual similarity to the victim as assessed by human perception. As a result, we sampled a total of 600 synthesized speech clips, with 200 samples capable of evading authentication for each of the speaker verification systems. These DeepFake audio samples were subsequently employed for AntiFake processing and evaluation. For each victim sample, we selected four target speakers distinct from the original victim. This selection included two speakers of the same gender as the victim and two speakers of a different gender, ensuring a diverse and representative set of speaker combinations.

Evaluation Metrics. Our primary focus is on three metrics that characterize the performance of AntiFake: authentication evasion reduction rate (AERR), perceptual speaker dissimilarity (PSD), and sound quality mean opinion score (MOS). Specifically, AERR is defined as the rate at which DeepFake samples can no longer bypass the corresponding SV system, given that all samples have already been filtered to bypass SV without AntiFake. As such, a higher AERR value represents a more effective system for preventing authentication evasion. The PSD is assessed by humans who evaluate the identity dissimilarity between synthesized samples and the original victim's speech sample. This human-rated metric is scored on a scale from 1 to 5, where 1 indicates the least dissimilar and 5 represents the most dissimilar. A higher PSD value signifies a greater divergence from the original speaker's voice and is therefore desired in our context. The MOS score in our experiments is derived from the well-established NISQA [35], a DNN-based speech assessment system that quantifies the overall quality and naturalness of speech on a scale from 1 to 5. This metric is used to measure the overall audio quality of the processed victim's sample, and therefore a higher value is desired. Empirically, a MOS of 3 or higher represents relatively good speech quality. For reference, the mean MOS for the TIMIT corpus is measured at 3.45±0.52.

Table 3: Overall performance of the two schemes in AntiFake.

	AEER			PSD	MOS
	GMM-UBM	ivector-PLDA	Azure		
Threshold-based	98.50%	98.75%	100%	4.85±0.33	3.44±0.61
Target-based	98.50%	99.50%	100%	4.88±0.47	3.32±0.58

6.6 Experimental Results

Overall Performance for Two Schemes. The overall performance of the two proposed schemes is summarized in Table 3. In terms of protection on speaker authentication systems (as measured by AEER), we observed that the protection rates for the three SV systems exceed 98%, demonstrating the effectiveness of AntiFake. Among the SV systems, AEER for Azure achieved the highest at 100%, while that of the GMM-UBM exhibited a relatively lower value. After manual examinations of these synthesized samples and evaluation results, we found that they exhibited high PSD scores as rated by humans (4.76±0.28), suggesting that the generated audio clips did not resemble the original speaker's voice. As such, we postulate that these samples may be considered "false positives" in SV systems, where they are mistakenly identified as originating from the enrolled speakers due to the lack of robustness in SV tasks.

Regarding the performance comparison of the two proposed schemes, no significant difference was observed in their protection efficacy. This can be attributed to the fact that both methods stem from the same underlying principle: sufficiently deviating the speaker embedding from its original value. However, the audio quality of speech samples processed using the threshold-based method appeared to be slightly better, potentially due to the larger solution space compared to targeting a specific speaker embedding. A more comprehensive breakdown of the results is presented in the subsequent ablation study as baselines.

Targeted Capability and Cross-gender Analysis. Although it is not our objective to transfer the speaker identity to another exact speaker, we investigated how well this can be done with AntiFake. To do so, we selected those generated with the target-based scheme and examined the perceptual speaker identity. The results indicated that while all of them were perceptually dissimilar to the victim's voice (therefore the protection is guaranteed), only 13.6% speech samples were marked as similar to the target speaker. This significant drop is within expectation for two reasons. First, the speaker boundaries for different synthesizers (or encoders) can be significantly different. As such, the targeted identity is difficult to transfer

Table 4: Ablation studies with different combinations of encoders.

	MOS	AdaptVC		SV2TTS		YourTTS		Tortoise		ElevenLabs	
		AERR	PSD	AERR	PSD	AERR	PSD	AERR	PSD	AERR	PSD
① AdaIN	3.84±0.22	99.6%	4.82±0.31	20.4%	2.42±0.80	0%	1.63±0.24	3.2%	1.42±0.20	0%	1.20±0.20
② GE2E	3.77±0.37	4.8%	1.30±0.20	100%	4.88±0.45	1.2%	1.1±0.18	20.6%	2.78±0.91	16.8%	1.25±0.56
③ H/ASP	3.80±0.20	5.6%	1.64±0.33	0%	1.20±0.20	100%	4.84±0.32	0%	1.45±0.39	13.6%	2.34±0.49
④ ViT	3.74±0.27	3.2%	1.34±0.21	1.2%	1.30±0.10	0%	1.48±0.37	100%	4.31±0.29	8.6%	2.34±0.24
①+②	3.62±0.33	100%	4.21±0.35	99.8%	4.33±0.48	3.6%	1.74±0.38	36.8%	3.21±0.49	88.6%	4.11±0.59
②+④	3.57±0.43	13.6$	2.01±0.46	100%	4.83±0.55	8.6%	1.96±0.28	79.8%	3.98±0.61	84.8%	3.78±0.66
①+②+③	3.46±0.34	100%	4.67±0.44	99.8%	4.22±0.57	100%	4.39±0.52	24.6%	2.28±0.45	90.6%	4.26±0.36
②+③+④	3.50±0.46	74.8%	3.88±0.62	100%	4.63±0.29	100%	4.16±0.22	98.6%	4.40±0.37	92.0%	4.40±0.30
①+②+③+④	3.37±0.60	99.4%	4.80±0.25	100%	4.89±0.42	99.8%	4.90±0.39	99.2%	4.73±0.36	97.7%	4.84±0.44

to unknown models. Second, our optimization strategy is designed to prioritize embedding deviation from the original one. Therefore, its ability to reach the target embedding is limited by design. We intentionally did this since strictly targeting a speaker embedding will significantly shrink the solution space, which however, may not be the optimal point that balances embedding deviation and perturbation magnitude. In cases where identity-targeting capabilities are desired, an additional penalty term can be added to restrict the discrepancies between optimized embedding and target embedding.

While the identity-targeting capability is limited by our design, it was interesting to find that the cross-gender success rate was as high as 82.4%. This means that, when the target is selected to be of a different gender than the victim, 82.4% synthesized speech goes to the gender of the target speaker. Such a high rate is not a coincidence; and we believe this is because even though the speaker boundaries for individual identities can be significantly different for various models, the acoustic features that characterize gender share more similarities across models [55]. As such, during the optimization towards the speaker of a different gender, those fundamental features are unavoidably altered and therefore lead to a high cross-gender success rate. From the perspective of protection, such phenomenon benefits AntiFake in that, when the synthesized audio is less likely to share the same gender as the victim, it becomes easier for victims to discern DeepFake speech since humans are more sensitive to distinguish speakers with different genders [21].

Run Time Analysis. In our experimental setup, we configured 1000 iterations of automatic optimization with the predetermined source and target speech samples. On average, the complete optimization process took 197.4 seconds to execute using an NVIDIA RTX 3090 GPU. Given the relatively short runtime, we regard AntiFake as a computationally efficient system that enhances usability.

6.7 Ablation Study

Impacts of Combining Different Encoders. In the previous experiments, we ensembled four speaker encoders to ensure robust protection against different models. To further investigate transferability across encoders, we analyzed the performance of various encoder combinations. The results are summarized in Table 4.

Overall, we found that the protection effectiveness (reflected in AEER and PSD) generally grows when the number of encoders increases. The protection reaches its best when all four encoders are combined for optimization. This is because the protection strength

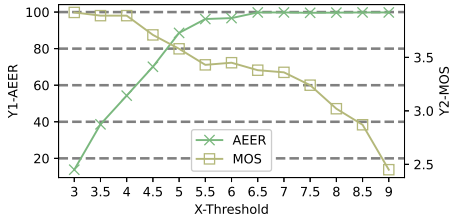

Figure 4: Measured AEER and MOS with varying thresholds.

is ultimately determined by the perturbations that can disrupt transferable acoustic features. When more feature extractors are employed, the extracted features will become more comprehensive across model architectures. This principle also explains why the transferability to commercial ElevenLabs is significantly lower when using individual encoders only. In this way, the optimization is equal to conducting a targeted adversarial attack against individual synthesizers as is done in [26], which is insufficient to form protection against unknown attacker models. Besides, we found that the length of the embedding does not explicitly affect the performance. This is because a longer embedding does not necessarily cover more comprehensive features that characterize the speaker's identity. Another observation is that the original quality generally degrades when ensembling more encoders. An intuition is that when combining more encoders to form a comprehensive set of acoustic features, it often requires more perturbations to sufficiently disrupt them, leading to relatively lower audio quality.

Impacts of Thresholds. We also studied how different thresholds can affect AntiFake in the threshold-based strategy, with results shown in Figure 4. We observed that when the threshold grows, the protection (AEER) increases since the optimization enforces a larger embedding deviation. As a side effect, however, the audio quality drops dramatically since it requires more perturbations.

Voice Features. AntiFake is built upon speaker embeddings to maximize protection. While such embeddings achieve state-of-the-art performance in extracting speaker features, they are latent space representations and are non-interpretable by nature. As such, we take a step further to understand the important voice features that need to be emphasized when developing protection. Specifically, we focused on four well-established features in the field of speaker recognition: Mel Frequency Cepstral Coefficient (MFCC) [50], Linear Frequency Cepstral Coefficient (LFCC) [67], Linear Predictive Coding (LPC) [51], and line spectral frequency (LSF) [31].

Table 5: Mean deviation factor of voice features.

	① AdaIN	② GE2E	③ H/ASP	④ ViT	Ensemble
MFCC	0.82	0.28	0.12	0.21	0.23
LFCC	0.55	1.28	0.66	3.24	0.83
LPC	0.28	0.65	1.63	6.33	1.22
LSF	0.46	0.17	0.37	0.28	0.20

The samples generated by AntiFake in previous experiments were utilized in this experiment. For each sample, we extracted the perturbations from the adversarial example ($\mathbf{x_U} + \delta_{\mathbf{x_U}}$) and original sample $\mathbf{x_U}$. The perturbations were then disrupted in terms of sequence order while their magnitudes were kept unchanged. In this way, we constructed control samples $\mathbf{x}^*_{\mathbf{U}}$ that were not protected but carried comparable noises. Subsequently, the aforementioned four features were extracted from both adversarial and control samples, denoted as f_{adv} and f_{ctrl} respectively. Intuitively, features that aid in protection will exhibit different values between these two sets. Based on this insight, we calculated the feature deviation factor as:

$$\text{Factor} = \left| \frac{(f_{adv} - f_U) - (f_{ctrl} - f_U)}{f_{ctrl} - f_U} \right|, \tag{11}$$

where f_U is the feature extracted from the original speech $\mathbf{x_U}$. As such, a higher value of this factor indicates a stronger correlation between the feature and protection.

This experiment was conducted on both individual encoders and their ensemble, with results summarized in Table 5. We observed that these encoders were associated with different features. For instance, the GE2E encoder manifested close relationships with LFCC features, while the ViT encoder mainly aligned with LPC features. Overall, the LPC features achieved the highest values across various encoders, indicating its potential to serve as a critical feature for protection. On the other hand, however, we did not identify any voice feature that was generalizable across all encoders. As such, explicitly focusing on specific voice features might only yield sub-optimal protection, highlighting the gap between voice features and speaker embeddings. While only the four most important voice features were examined, this insight could inspire future work aimed at exploring other features and their optimal combination toward improved protection.

6.8 Evaluation against Adaptive Attackers

We also evaluated AntiFake in the context of adaptive attackers. In this scenario, attackers aware of the existence of perturbations will attempt to remove them and proceed with synthesis with sanitized speech audio. We consider two categories of strategies, where the attacker may (1) employ a set of transformation operations to invalidate perturbations, or (2) use optimization to remove perturbations.

Speech Sample Transformation. For evaluation, we follow WaveGuard [27] which is a recent work proposed to undermine adversarial perturbations via signal processing techniques. Specifically, we implemented four audio transformations: (1) quantization-dequantization, (2) down-sampling and up-sampling, (3) frequency filtering, and (4) mel-spectrogram extraction and inversion. We followed the same setup as the original implementation and used an identical set of parameters to ensure consistency with prior work and transparent comparison. The transformations are summarized as follows, and for more details, we refer to the original paper [27].

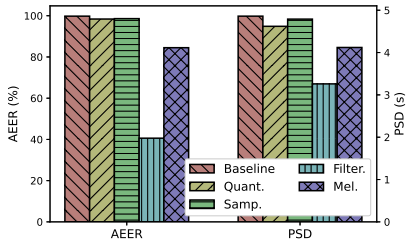

Figure 5: AntiFake performance against adaptive attackers with transformation-based strategies.

Quantization-Dequantization. We quantized the waveform audio into 8 bits and reconstructed it to approximate the original data.

Down-sampling and Up-sampling. We downsampled the original waveform (16 kHz) to lower frequencies (12kHz, 10kHz, and 8kHz) that are suggested to be the most optimal defensive sampling rates, and upsampled it via interpolation back to the original sample rate.

Frequency Filtering. Frequency filtering attenuates the signal above and below certain thresholds with high/low-shelf filters. In our experiment, we first computed the spectral centroid of each waveform, then applied a negative gain of 30 on the amplitude of frequencies above 1.5 times the centroid and below 0.1 times the centroid.

Mel-spectrogram Extraction and Inversion. The sample was converted to mel-spectrogram and then converted back to waveform.

Optimization-based Perturbation Removal. As perturbations are optimized in AntiFake, the attacker can also adopt a similar strategy to remove them. However, due to the lack of the original speaker embedding, the attacker is not able to follow the same optimization process (Section 5.4) to reconstruct the original speech directly. Therefore, the attacker has to rely on SV models or humans to determine the optimization direction and the convergence state. Two types of adversarial feedback are incorporated: (1) the speech quality that needs to be maximized (i.e., minimize distortions), and (2) the extent to which the DeepFake audio can pass SV systems (measured by SV outputs) or deceive humans (quantified as PSD).

For evaluation, the MOS score paired with human perception was used to assess speech quality. The attacker is also involved in using human perception to rate the PSD score judging the effectiveness of DeepFake samples in deceiving humans. The ivector-PLDA was selected as the target SV system, and it was set up in two configurations: one with the confidence score exposed to the attacker while the other does not, and the score was termed to be maximized. Due to the attacker's realistic inability to access gradients within target SV systems and the non-differentiable nature of the human feedback involved, we built upon a search-based optimization framework [49]. Considering the significant computational costs and manual labor, 100 samples were used in this experiment.

Evaluation Results. The results for the transformation-based approach are summarized in Figure 5. Overall, the protection remains relatively robust for two reasons. First, the perturbations are not entirely removed after processing. The defensive nature of such transformations necessitates restoring the original waveform with minimal harm to the audio quality. However, our optimization

driven by psychoacoustic principles closely ties the perturbations to the speech waveform. As such, it becomes difficult for processing techniques to eliminate perturbations without significantly harming the audio. Second, our perturbations are optimized specifically for disrupting features that characterize speaker embeddings. While the full set can guarantee high performance, those that persist after processing can still shift the speaker embedding. This is a fundamental difference from traditional adversarial audio where the perturbations are designed to reach targeted phrases or classification labels, and the narrow solution space renders their perturbations relatively "fragile" to be broken by transformations.

Among the four transformations, we found that frequency filtering degraded the AntiFake protection the most, due to the fact the employed filters removed more perturbations compared to other transformations. However, upon manual hearing on the resulting speech samples derived from frequency filtering, we found the speech content suffers non-trivial alternation. This further facilitates that the synthesized audio sounds blurred and unnatural to humans, and simultaneously none of them can pass the authentication systems. In this context, the DeepFake audio is less likely to cause harm. To sum up, this reflects a fundamental trade-off on the attacker side, where stronger filtering can break the protection provided by AntiFake, yet the significantly undermined audio will result in low-quality DeepFake detrimental to the malicious goals.

On the other hand, the optimization-based approach exhibited different challenges. Even with the confidence score exposed to the attacker, we found that only three DeepFake samples passed the SV system, and their number of iterations needed were 8615, 7967, and 9034, respectively. Besides, six samples attained a PSD of higher than 3. Other than that, the rest optimization trials failed after reaching the 10K iteration limit. When the confidence score is not accessible, only two samples achieved a PSD higher than 3 and none of the resulting DeepFake speech can bypass the SV system within the query limit. This is because without the score, the query feedback is merely a binary of either "Accept" or "Reject", rendering the optimization much less effective. These results showed the feasibility of creating usable DeepFake samples by using adaptive optimization. While only a small number of samples achieved the adversarial goal, it does not imply the attack is less powerful. Conversely, this strategy provides finer-grained attack capabilities, and the perturbations could always be found given unlimited queries. In practical scenarios, however, the large number of queries poses significant barriers for the attacker in terms of both computation and human efforts. At the later stage of optimization, we found that human-rated terms could sometimes contradict the values when rated normally. This could be attributed to the so-called semantic satiation, where repeated listening queries might cause cognitive degradation. These results also provide valuable insights for defensive measures, which are discussed further in Section 7.

6.9 AntiFake Usability Test

A key objective of AntiFake is to make protection accessible to the public, for which reason usability is a crucial desired property. Therefore, we also conducted usability tests followed by surveys to evaluate practical usage. The experiments and surveys involved were approved by the local Institutional Review Board (IRB).

Participants Recruitment. We first surveyed a larger group of volunteers for demographic information and subsequently sampled a total of 24 participants. The participant group consisted of 13 (54.2%) males and 11 (45.8%) females; 15 (62.5%) aged 18-29, 5 (20.8%) aged 30-50, and 4 (16.7%) aged over 50. Regarding educational levels, we selected participants to be evenly distributed across four categories: "no high school diploma or equivalent," "high school diploma or equivalent," "college degree," and "graduate or professional degree." As such, we aimed to recruit participants with diverse backgrounds to form a representative population of AntiFake users.

Usability Test Methodology. We provided each participant with a set of source speech samples and asked them to follow the instructions of AntiFake. During each round of experiment, they were tasked to transcribe the processed sample before listening to the original speech. Participants were given unlimited time and attempt to experience the system. Upon completing the experiments, they were asked to fill out a survey incorporating the standard System Usability Scale (SUS) questionnaire, assess the speech quality of the protected samples, and provide open-ended feedback. In terms of speech quality assessment, the transcription correctness was analyzed to serve as an indicator of speech clarity and quality. Besides, the participants were asked to rate an integer score from 1 to 5, with 1 representing the worst and 5 indicating the best quality. To mitigate desirability bias, the research goal was hidden from the participants. The complete survey can be found in Appendix A.

Usability Test Results. The SUS scores for individual responses were calculated, yielding an average of 87.60 (±4.81) across all responses. In reference to the SUS curve and percentile ranks [41], AntiFake achieved an above-average SUS score of 68 at 50% percentile. Besides, the human-perceived speech quality scores were averaged at 3.54 ± 0.59. This result aligns with our measured MOS scores calculated by the NISQA model, highlighting the relatively good speech quality of the processed samples. We also manually verified the transcriptions from the participants, and found that all of them aligned with ground-truth contents. As such, it provides evidence that the protective processing of AntiFake preserves the original speech quality and does not compromise the normal human comprehension of these samples.

Regarding the open-ended feedback, the majority of participants (n=17, 70.8%) expressed amazement at the superior performance of speech synthesizers and agreed that preventative measures should be in place for mitigation. Many of them (n=8, 33.3%) characterized the system as "easy to use". Some participants (n=4, 16.7%) expressed a desire for features enabling simplified user involvement, suggesting a "one-click functionality" for immediate output.

7 DISCUSSION AND LIMITATIONS

Real-world Limitations. The effectiveness of AntiFake is dependent on users consistently applying it to their audio before public release. In reality, however, users may not always have control over all instances of their voice recordings. While it might not be easy for the attacker to link samples to the target individual, such unprocessed samples could create opportunities for attackers to undermine the protection. To enhance overall security, it is better for users to adopt a multi-faceted approach to safeguard their online speech data. This could include careful curation of their

digital presence, untagging or removing themselves from publicly accessible recordings, and taking advantage of privacy regulations to limit the availability of their voice samples online.

On the other hand, AntiFake's protection could be compromised by future techniques, which is known as the challenge of being *future-proof* [43]. For instance, imminent synthesizers might not use speaker embeddings, or advanced audio purification techniques might emerge to remove perturbations even without any prior knowledge of the original speaker or samples. This further highlights the need for continued research into proactive privacy protection strategies, especially with an emphasis on sustained protection.

Overlapped Encoders. The encoders of the evaluated open-source synthesizers are overlapped with those integrated within AntiFake. However, this does not undermine our black-box settings. The key insight of AntiFake is that encoder is a general architecture within synthesizers and shares similarities due to their common goal of robust embedding extraction. Therefore, AntiFake relies on transferability to disrupt black-box synthesizers of attackers. We experimentally validated this point from two aspects. First, AntiFake was evaluated against the commercial ElevenLabs, which is a pure black-box system employing unknown encoders. With our designed encoder ensemble, AntiFake was shown to provide robust protection. Second, we further delved into the impacts of encoders with a fine-grained ablation study. As suggested by Table 4, different levels of transferability exist even when entirely different encoders were used for optimization and synthesis. Besides, the increased number of ensembled encoders generally led to augmented protection, and leveraging the complete set of the four encoders achieved the best protection even against black-box products. We believe such an initial clue of transferability is a crucial insight that can inspire future studies, which can further investigate the optimal combination across a wide range of encoders.

Adaptive Attackers. In this study, we also considered adaptive attackers leveraging speech transformation and optimization-based perturbation removal as two key strategies. Through the experiments, they exhibited different characteristics. For the transformation-based approach, we found that certain processing methods (e.g., filtering) can indeed degrade the protection, however, the strong filtering also undermines the speech sample and therefore leads to low-quality DeepFake samples. As such, it is a trade-off that the attacker has to make between the fidelity of DeepFake speech and the strength to remove perturbations. On the other hand, the optimization-based method provides finer-grained capabilities to find the perturbations; however, it requires a large number of queries to SV models or humans that cause significant yet realistic barriers for the attacker. Inspired by these results, some measures could be taken to counteract such adaptive attackers. For example, targeting the large number of queries needed, rate-limiting the queries to SV systems or capping unsuccessful attempts could be leveraged to mitigate threats. Moreover, removing the user's access to confidence scores could further hinder the attacker's attempts.

Multiple and Longer Samples. In practical scenarios, the user may seek to use AntiFake to protect longer or even multiple samples. AntiFake is not principally limited by the sample length. This is because samples with different lengths are mapped to fixed-length speaker embeddings. Therefore, the protection level determined by the embedding deviation is not affected by audio lengths. Furthermore, multiple speech samples can be batch-processed with one target embedding. While such samples may expose more information to the attacker, the consistent embedding of the targeted persona ensures that the advantage given to the attacker is limited.

This naturally leads to the question of how much protection is sufficient. There are two factors, the amount of perturbations and the costs of removing them. For perturbation, there is a delicate balance between protection strength and sample quality. The more perturbations, the harder it is to reverse, but likely worse quality. Furthermore, perturbation magnitude is closely associated with the acceptable risk and attacker cost. Such costs include both human capital and computational resources, as revealed in the study against adaptive attackers. For countermeasures, traditional security techniques like rate limiting can further improve the asymmetry.

Desirability Bias. In designing the human studies and associated surveys, we have taken measures to minimize ambiguity and communication inefficiency. For example, rather than requiring participants to comprehend technical terms such as "adversarial examples" and intricate internal workings, we designed end-to-end human evaluation of AntiFake as a system. Additionally, we made efforts to mitigate desirability bias by incorporating both positive and negative questions in the survey and hiding the study goal from the participants. Despite these efforts, however, unobserved desirability bias from the researcher's perspective may still persist. For instance, participants might deduce that they are evaluating the efficacy of a novel tool developed by researchers. As a result, some participants might be more likely to provide positive feedback which they believe is advantageous and appealing to the researcher.

Ethical Considerations. We care deeply about the security of our society and have strived to address the potential ethical considerations associated with our work. First, all of the aforementioned user studies strictly follow the protocols approved by the local IRB. Throughout the process, the participants were made aware that the information conveyed within the speech samples was not real and not representative of actual events or individuals. Second, all generated DeepFake speech samples, particularly those exhibiting high levels of realism sufficient to bypass authentication systems and deceive human perception, are not used outside of the study and have been deprecated following the conclusion of the research.

8 CONCLUSION

In this work, we propose AntiFake, a preventative defense against DeepFake audio threats. complementary to existing detection methods, AntiFake leverages adversarial perturbations to hinder unauthorized speech synthesis. To improve usability and accessibility to diverse populations, AntiFake is designed as a human-in-the-loop system involving minimal human efforts. The efficacy of AntiFake is evaluated with state-of-the-art synthesizers, and the usability is validated with usability tests involving human participants.

ACKNOWLEDGMENT

We thank the reviewers for their valuable feedback. This work is supported in part by the NSF (CNS-1916926, CNS-2038995, CNS-2154930, CNS-2238635), and ARO (W911NF2010141).

REFERENCES

[1] Hadi Abdullah et al. 2019. Practical Hidden Voice Attacks against Speech and Speaker Recognition Systems. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019.*

[2] Hadi Abdullah et al. 2021. Hear "No Eil", See "Kenansville"*: Efficient and transferable black-box attacks on speech recognition and voice identification systems. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 712–729.

[3] Muhammad Ejaz Ahmed et al. 2020. Void: A fast and light voice liveness detection system. In *29th USENIX Security Symposium (USENIX Security 20)*. 2685–2702.

[4] James Betker. 2022. TorToiSe TTS. https://github.com/neonbjb/tortoise-tts.

[5] Logan Blue et al. 2022. Who Are You (I Really Wanna Know)? Detecting Audio DeepFakes Through Vocal Tract Reconstruction. In *31st USENIX Security Symposium, USENIX Security 2022*. USENIX Association, 2691–2708.

[6] Douglas S. Brungart. 2001. Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America* 109, 3 (03 2001), 1101–1109.

[7] Nicholas Carlini et al. 2016. Hidden voice commands. In *25th USENIX Security Symposium (USENIX Security 16)*. 513–530.

[8] Nicholas Carlini et al. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 1–7.

[9] Edresson Casanova et al. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*. PMLR, 2709–2720.

[10] Guangke Chen et al. 2021. Who is real bob? adversarial attacks on speaker recognition systems. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE.

[11] Tao Chen et al. 2020. Metamorph: Injecting Inaudible Commands into Over-the-air Voice Controlled Systems. In *27th Annual Network and Distributed System Security Symposium, NDSS 2020, San Diego, California, USA, February 23-26, 2020.*

[12] Valeriia Cherepanova et al. 2021. LowKey: Leveraging Adversarial Attacks to Protect Social Media Users from Facial Recognition. In *9th International Conference on Learning Representations, ICLR 2021, Austria, May 3-7, 2021.*

[13] Ju-Chieh Chou et al. 2019. One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization. In *20th Annual Conference of the International Speech Communication Association*. ISCA, 664–668.

[14] Graham Cluley. 2022. Deepfaking crooks seek remote-working jobs to gain access to sensitive data. https://grahamcluley.com/deepfaking-crooks-seek-remote-working-jobs-to-gain-access-to-sensitive-data/.

[15] Joseph Cox. 2023. How I Broke Into a Bank Account With an AI-Generated Voice. https://www.vice.com/en/article/dy7axa/how-i-broke-into-a-bank-account-with-an-ai-generated-voice.

[16] Najim Dehak et al. 2010. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 4 (2010), 788–798.

[17] Alexey Dosovitskiy et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021.*

[18] ElevenLabs. 2023. Prime Voice AI. https://beta.elevenlabs.io/.

[19] John S Garofolo et al. 1993. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report* (1993).

[20] Ian J. Goodfellow et al. 2015. Explaining and Harnessing Adversarial Examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.*

[21] Antonio Guerrieri et al. 2022. Gender identification in a two-level hierarchical speech emotion recognition system for an Italian Social Robot. *Sensors* 22, 5 (2022), 1714.

[22] Hanqing Guo et al. 2022. Specpatch: Human-in-the-loop adversarial audio spectrogram patch attack on speech recognition. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 1353–1366.

[23] Hee Soo Heo et al. 2020. Clova baseline system for the voxceleb speaker recognition challenge 2020. *arXiv preprint arXiv:2009.14153* (2020).

[24] Kenji Homma et al. 2009. Ossicular resonance modes of the human middle ear for bone and air conduction. *The Journal of the Acoustical Society of America* 125, 2 (2009), 968–979.

[25] HSBC. 2016. How do I sign up for Voice ID? https://www.hsbc.co.uk/ways-to-bank/phone-banking/.

[26] Chien-yu Huang et al. 2021. Defending your voice: Adversarial attack on voice conversion. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE.

[27] Shehzeen Hussain et al. 2021. {WaveGuard}: Understanding and Mitigating Audio Adversarial Examples. In *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*. USENIX Association, 2273–2290.

[28] Won Jang et al. 2021. UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association*. ISCA, 2207–2211.

[29] Corentin Jemine. 2019. Real-time-voice-cloning. https://github.com/CorentinJ/Real-Time-Voice-Cloning. *University of Liége, Liége, Belgium* (2019).

[30] Ye Jia et al. 2018. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*. 4485–4495.

[31] Tomi Kinnunen et al. 2010. An overview of text-independent speaker recognition: From features to supervectors. *Speech communication* 52, 1 (2010).

[32] Jungil Kong et al. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems* 33 (2020), 17022–17033.

[33] Felix Kreuk et al. 2018. Fooling end-to-end speaker verification with adversarial examples. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 1962–1966.

[34] Zhuohang Li et al. 2020. Practical adversarial attacks against speaker recognition systems. In *Proceedings of the 21st international workshop on mobile computing systems and applications*.

[35] Gabriel Mittag et al. 2021. NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association*. ISCA, 2127–2131.

[36] Yishuang Ning et al. 2019. A Review of Deep Learning Based Speech Synthesis. *Applied Sciences* 9, 19 (2019).

[37] Vassil Panayotov et al. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE.

[38] Daniel Povey et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

[39] Kaizhi Qian et al. 2019. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*. PMLR, 5210–5219.

[40] Douglas A Reynolds et al. 2000. Speaker verification using adapted Gaussian mixture models. *Digital signal processing* 10, 1-3 (2000), 19–41.

[41] Jeff Sauro et al. 2016. *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.

[42] Lea Schönherr et al. 2019. Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding. In *26th Annual Network and Distributed System Security Symposium (NDSS)*.

[43] Shawn Shan et al. 2020. Fawkes: Protecting Privacy against Unauthorized Deep Learning Models. In *29th USENIX Security Symposium, USENIX Security 2020.*

[44] Jiacheng Shang et al. 2018. Defending against voice spoofing: A robust software-based liveness detection system. In *2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 28–36.

[45] Jonathan Shen et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4779–4783.

[46] Sayaka Shiota et al. 2015. Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification. In *Sixteenth annual conference of the international speech communication association.*

[47] Catherine Stupp. 2019. Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

[48] Yôiti Suzuki et al. 2004. Equal-loudness-level contours for pure tones. *The Journal of the Acoustical Society of America* 116, 2 (08 2004).

[49] Rohan Taori et al. 2019. Targeted adversarial examples for black box audio systems. In *2019 IEEE security and privacy workshops (SPW)*. IEEE, 15–20.

[50] Vibha Tiwari. 2010. MFCC and its applications in speaker recognition. *International journal on emerging technologies* 1, 1 (2010), 19–22.

[51] Satyam P Todkar et al. 2018. Speaker recognition techniques: A review. In *2018 3rd International Conference for Convergence in Technology (I2CT)*. IEEE, 1–5.

[52] Uberduck. 2023. Text to Voice. https://app.uberduck.ai/voice-to-voice. (2023).

[53] Omkarprasad S Vaidya et al. 2006. Analytic hierarchy process: An overview of applications. *European Journal of operational research* 169, 1 (2006), 1–29.

[54] Aäron van den Oord et al. 2016. WaveNet: A Generative Model for Raw Audio. In *The 9th ISCA Speech Synthesis Workshop, September 2016*. ISCA, 125.

[55] Rivarol Vergin et al. 1996. Robust gender-dependent acoustic-phonetic modelling in continuous speech recognition based on a new automatic male/female classification. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, Vol. 2. IEEE.

[56] James Vincent. 2023. 4chan users embrace AI voice clone tool to generate celebrity hatespeech. https://www.theverge.com/2023/1/31/23579289/ai-voice-clone-deepfake-abuse-4chan-elevenlabs.

[57] Tarun Wadhwa. 2015. Wells Fargo Wants To Let You Make Million-Dollar Wire Transactions With Your Face And Voice. https://www.forbes.com/sites/tarunwadhwa/2015/11/03/why-wells-fargo-wants-to-let-you-make-million-dollar-wire-transactions-with-your-face-and-voice/.

[58] Li Wan et al. 2018. Generalized End-to-End Loss for Speaker Verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, April 2018*. IEEE, 4879–4883.

[59] Steven H Weinberger et al. 2011. The Speech Accent Archive: towards a typology of English accents. In *Corpus-based studies in language use, language learning, and language documentation*. Brill, 265–281.

[60] Emily Wenger et al. 2021. Hello, It's Me: Deep Learning-based Speech Synthesis Attacks in the Real World. In *Proceedings of the 2021 ACM SIGSAC Conference on*

Computer and Communications Security. 235–251.
[61] Junichi Yamagishi et al. 2019. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)* (2019).
[62] Zhiyuan Yu et al. 2021. Security and privacy in the emerging cyber-physical world: A survey. *IEEE Communications Surveys & Tutorials* 23, 3 (2021), 1879–1919.
[63] Zhiyuan Yu et al. 2023. {SMACK}: Semantically Meaningful Adversarial Audio Attack. In *32nd USENIX Security Symposium (USENIX Security 23)*. 3799–3816.
[64] Xuejing Yuan et al. 2018. Commandersong: A systematic approach for practical adversarial voice recognition. In *27th USENIX Security Symposium (USENIX Security 18)*. 49–64.
[65] Anna Zhadan. 2023. Emma Watson reads Mein Kampf while Biden announces invasion of Russia in latest AI voice clone abuse. https://cybernews.com/news/ai-voice-clone-misuse/.
[66] Linghan Zhang et al. 2017. Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 57–71.
[67] Xinhui Zhou et al. 2011. Linear versus mel frequency cepstral coefficients for speaker recognition. In *2011 IEEE workshop on automatic speech recognition & understanding*. IEEE, 559–564.

A SURVEY QUESTIONS FOR USABILITY TEST

(1) What is your age group?
- 18-29
- 30-50
- Over 50

(2) What is your gender?
- Male
- Female
- Prefer not to say

(3) What is your highest level of education?
- No high school diploma or equivalent
- High school diploma or equivalent (e.g., GED)
- College degree (e.g., associate or bachelor's degree)
- Graduate or professional degree (e.g., MS, PhD, MD, JD)

(4) I think that I would like to use this system frequently.
(5) I found the system unnecessarily complex.
(6) I thought the system was easy to use.
(7) I think that I would need the support of a technical person to be able to use this system.
(8) I found the various functions in this system were well integrated.
(9) I thought there was too much inconsistency in this system.
(10) I would imagine that most people would learn to use this system very quickly.
(11) I found the system very cumbersome to use.
(12) I felt very confident using the system.
(13) I needed to learn a lot of things before I could get going with this system.
(14) Please provide your perceived transcriptions of the processed speech samples in the following.
(15) On a scale from 1 to 5, please rate your perception of the audio quality of the processed speech samples. Note that 1 represents the worst quality while 5 indicates the best.
(16) Please share your thoughts on any aspects of AntiFake, including its advantages, disadvantages, or any other observations. Feel free to comment on its performance, usability, and any suggestions for improvement.

Table 6: Sentences that carry various malicious intents when used for DeepFake speech synthesis.

Malicious Intents	Sentences
Conduct Financial Scam	I need money for an emergent surgery, please transfer me a thousand dollars.
	I need a grand for an emergency medical bill.
	I'm in trouble in a foreign country, can you lend me some money?
	I need a hundred for a last minute flight. Can you help me out?
	I lost my wallet, can you lend me a hundred?
	You still owe me a thousand dollars, can you send it to my new account?
	Your payment did not work, can you try again to my new account?
Compromise Safety and Privacy	I have arrived at your house, please open the door.
	Help me recall my cell phone password.
	I highly recommend you install this free software.
	I have sent you the contract, please click the link to sign.
	Please forward me the 2FA code for account login.
	I forgot my SSN number, can you send it to me?
Spread Hate speech or Misinformation	There has been an emergency, all residents must evacuate immediately.
	A magnitude 8 earthquake just hit California.
	A major chemical spill happened in New York.
	Cancer is caused by vaccination.
	A blood test for coronavirus costs three thousand dollars.
	Smoking can help prevent the coronavirus.
	Cocaine is a cure for COVID 19.
	I wish you were never born.
Bypass Voice-based Authentication	My voice is my password.
	My bank account is secured by my voiceprint.
	I confirm the transaction.