ClearMask: Noise-Free and Naturalness-Preserving Protection Against Voice Deepfake Attacks

Yuanda Wang
Michigan State University
East Lansing, Michigan, USA
wangy208@msu.edu

Bocheng Chen
Michigan State University
East Lansing, Michigan, USA
chenboc1@msu.edu

Hanqing Guo
University of Hawaii at M?noa
Honolulu, Hawaii, USA
guohanqi@hawaii.edu

Guangjing Wang
University of South Florida
Tampa, Florida, USA
guangjingwang@usf.edu

Weikang Ding
Michigan State University
East Lansing, Michigan, USA
dingweik@msu.edu

Qiben Yan
Michigan State University
East Lansing, Michigan, USA
qyan@msu.edu

Abstract

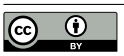
Voice deepfake attacks, which artificially impersonate human speech for malicious purposes, have emerged as a severe threat. Existing defenses typically inject noise into human speech to compromise voice encoders in speech synthesis models. However, these methods degrade audio quality and require prior knowledge of the attack approaches, limiting their effectiveness is diverse scenarios. Moreover, real-time audios, such as speech in virtual meetings and voice messages, are still exposed to voice deepfake threats. To overcome these limitations, we propose ClearMask, a noise-free defense mechanism against voice deepfake attacks. Unlike traditional approaches, ClearMask modifies the audio mel-spectrogram by selectively filtering certain frequencies, inducing a transferable voice feature loss without injecting noise. We then apply audio style transfer to further deceive voice decoders while preserving perceived sound quality. Finally, optimized reverberation is introduced to disrupt the output of voice generation models without affecting the naturalness of the speech. Additionally, we develop ClearMask, named LiveMask, to protect streaming speech in real-time through universal frequency filter and reverberation generator. Our experimental results show that ClearMask and LiveMask effectively prevent voice deepfake attacks from deceiving speaker verification models and human listeners, even for unseen voice synthesis models and black-box API services. Furthermore, ClearMask demonstrates resilience against adaptive attackers who attempt to recover the original audio signal from the protected speech samples.

CCS Concepts

• **Security and privacy**; • **Computing methodologies** → **Machine learning**;

Keywords

Voice Synthesis, Adversarial Machine Learning, Privacy



ASIA CCS '25, Hanoi, Vietnam

ACM ISBN 979-8-4007-1410-8/25/08
https://doi.org/10.1145/3708821.3733881

ACM Reference Format:
Yuanda Wang, Bocheng Chen, Hanqing Guo, Guangjing Wang, Weikang Ding, and Qiben Yan. 2025. ClearMask: Noise-Free and Naturalness-Preserving Protection Against Voice Deepfake Attacks. In *ACM Asia Conference on Computer and Communications Security (ASIA CCS '25), August 25–29, 2025, Hanoi, Vietnam.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3708821.3733881

1 Introduction

With the rapid advancement of deep learning, voice synthesis models have become increasingly powerful, producing speech sounds highly natural and lifelike [3, 5]. Building on it, state-of-the-art voice synthesis models can create convincing speech content using a short speech sample as a reference, replicating not only the speaker's voice but also their prosody and rhythm to make the output indistinguishable from real human speech [48]. However, these advancements also lead to threats when generated speech is misused, known as voice deepfake attacks [50]. In 2023, fraudsters successfully bypassed bank authentication systems using synthetic voices [4], and even orchestrated a fake kidnapping by synthesizing a girl's voice [6]. In another instance, criminals deceived a company into transferring $200,000 using fake speech [2]. Some individuals have exploited this technology to generate hate speech using the voices of celebrities, causing significant reputational harm [53]. Additionally, synthetic voices can be used to compromise voice assistants during user verification processes or to execute malicious voice commands [49]. These examples underscore the risks of synthetic speech being used to deceive humans or automatic speaker verification (ASV) systems. Unfortunately, preventing the misuse of synthetic speech by restricting speech synthesis models remains a substantial challenge. As a result, exposing unprotected voices on public media platforms raises serious threats to user security and privacy.

One common defense against artificially generated speech is liveness detection [10], which identifies unnatural speech not originate from human vocal tracts. However, this approach is limited as it cannot prevent attacks designed to deceive human listeners. To address the limitation, Attack-VC [24] mitigates voice deepfake attacks by masking speech audio before it is uploaded to public social media platforms. By injecting carefully optimized perturbations into audio signals, Attack-VC prevents attackers from generating convincing

Table 1: Comparison of ClearMask with existing defenses.

Defenses	Effectiveness	Transferability	High-quality	Real-time
Attack-VC [24]	Medium	Low	Low	Low
SampleMask [32]	Low	Medium	Medium	High
VSMask [46]	High	Low	Low	High
AntiFake [52]	High	High	Medium	Low
ClearMask	**High**	**High**	**High**	**High**

speech samples capable of deceiving ASV systems or human listeners. Despite its effectiveness in specific scenarios, achieving a comprehensive defense in diverse real-world applications remains challenging. These challenges can be categorized into four key aspects. First, achieving a high success rate is critical for all defense mechanisms. Even a small chance of successful attacks can compromise the defense, leading to severe and unacceptable damages. Second, defenders often lack prior knowledge about the specific voice synthesis models that attackers employ. Therefore, existing white-box defenses cannot be generalized to protect against various voice synthesis models. Third, the protected speech audio must retain clarity and intelligibility to ensure its usability in practical applications, such as videos or voice messages. Fourth, safeguarding real-time speech audio in instant communication applications, e.g., FaceTime [1] and online meetings, is essential. These scenarios, however, require protection mechanisms to function with minimal latency, which presents a technical challenge.

Consequently, we identify four essential features for effective protection against voice deepfake attacks: *effectiveness, transferability, high-quality*, and *real-time*, corresponding to the four challenges outlined above. In Table 1, we evaluate existing defenses across these dimensions. However, while individual defenses demonstrate unique strengths, none of them satisfy all four requirements. Attack-VC [24] and VSMask [46] are white-box defenses with limited transferability. SampleMask [32] requires querying a specific model, and demonstrates limited effectiveness. While AntiFake [52] exhibits strong effectiveness and transferability, it suffers from poor audio quality and lacks real-time feasibility. To overcome these challenges, we propose ClearMask, a noise-free defense against voice deepfake attacks. While ensuring *effectiveness*, we adopt an ensemble encoder approach to enhance the *transferability*. Moreover, unlike traditional methods which inject noise to mask voice samples, ClearMask leverages multiple natural sound effects to generate *high-quality* speech. Specifically, it first modifies the input mel-spectrogram of voice synthesis models by filtering out specific frequencies in speech spectrograms. Next, ClearMask employs audio style transfer to obscure distinctive voice features, effectively misleading deepfake voice generation. Finally, it introduces a well-optimized room impulse response (RIR) to create unique reverberation effects that further enhance the protection. In addition, we design LiveMask, a *real-time* mode of ClearMask designed for online speech protection. LiveMask employs universal frequency filter and reverberation sound effect to provide immediate protection for instant communication scenarios.

In the evaluation, we test ClearMask and LiveMask on both open-source and commercial voice synthesis platforms. The results demonstrate that ClearMask and LiveMask effectively prevent unauthorized voice synthesis in both offline and online scenarios. With this protection, the generated deepfake voices fail to deceive ASV systems or human listeners. Furthermore, ClearMask demonstrates robustness against adaptive attackers with varying capabilities who attempt to remove the reverberation, effectively maintaining its defensive performance. Overall, our contributions are summarized as follows:

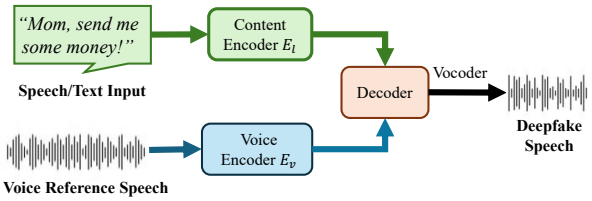


Figure 1: A general framework of voice synthesis models.

- We introduce ClearMask, a noise-free protection mechanism against malicious voice cloning. ClearMask employs spectrogram masking to modify the mel-spectrogram. In addition, we utilize audio style transfer and artificial reverberation to further obscure distinctive voice features.
- We propose LiveMask, a real-time mode of ClearMask designed for immediate protection. It applies a general frequency filter and universal reverberation generation to protect streaming online speech with millisecond-level latency.
- We provide a comprehensive system design to optimize the three stages of ClearMask. Through a surrogate ensemble encoder, we strike a balance among effectiveness, transferability and audio quality.
- We comprehensively evaluate the protection performance of ClearMask and LiveMask on five voice synthesis models including commercial APIs. Our results demonstrate that ClearMask achieves effective, transferable, high-quality, and real-time protection in various scenarios.

2 Preliminaries

2.1 Voice Deepfake Attacks and Defenses

Voice synthesis models produce artificial speech by integrating linguistic content derived from source speech or textual input with voice features extracted from a reference speech sample. Voice synthesis approaches are typically categorized into Voice Conversion (VC), which modifies the voice of a given speech sample to mimic a target speaker [18, 19, 38], and Text-to-Speech (TTS), which generates speech from text using the target speaker's voice [13]. Fig. 1 illustrates a general framework of voice synthesis, where content and voice features are encoded into embeddings, decoded to produce a mel-spectrogram, and converted into audible waveforms. VC usually creates more expressive results by retaining human emotion and intonation, while advanced TTS models also achieve high realism with style embedding [48]. To prevent unauthorized voice synthesis, users can employ adversarial examples to compromise the voice encoder module E_v, leading to unqualified synthetic voice samples. The optimization process of this adversarial speech is:

$$arg\max_{\delta} \|E_v(\boldsymbol{x}_r + \delta) - E_v(\boldsymbol{x}_r)\|_2, \quad \textbf{s.t.}\ \|\delta\|_\infty < \epsilon, \tag{1}$$

where $\boldsymbol{x}_r$ is the victim's speech sample and δ is perturbation signal. However, this approach heavily depends on gradient backpropagation from a single encoder model, which limits its transferability in black-box scenarios. Additionally, the perturbations δ

degrade the audio quality of the adversarial speech samples. Therefore, more advanced methods are needed to effectively protect speech against various attack techniques while preserving audio quality and naturalness.

2.2 Audio Style Transfer

Inspired by image style transfer [21], audio style transfer has emerged with the goal of modifying the texture of sound, such as the timbre of a musical instrument [22, 42]. Both audio style transfer and voice conversion focus on embedding features of one audio sample into another to synthesize a new audio. An audio style transfer model generally includes a style extraction module $\boldsymbol{E}_s$ and a style synthesis module $\boldsymbol{S}$, and can be mathematically represented as follows:

$$\boldsymbol{x}_t' = \boldsymbol{S}(\boldsymbol{x}_t, V_s), \quad \textbf{s.t. } V_s = \boldsymbol{E}_s(\boldsymbol{y}_t), \tag{2}$$

where $\boldsymbol{x}_t$ is the source audio and $\boldsymbol{y}_t$ is the style reference audio. Although audio style transfer does not affect the speech content features, the difference in sound texture could mislead the voice encoder in voice synthesis models and damage the extracted voice embedding vectors.

2.3 Reverberation in Adversarial Examples

Reverberation is a phenomenon caused by multi-path effects in the physical world. The reverberation can be quantified using RIR, defined as the response signal after a pulse is played in a given environment. We can simulate reverberation by convolving an RIR with an original audio $\boldsymbol{x}_t$:

$$\boldsymbol{x}_t^R = \boldsymbol{x}_t * h, \tag{3}$$

where h is the reversed RIR signal, and $\boldsymbol{x}_t^R$ is the audio with reverberation effect.

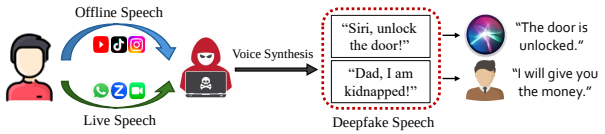
While reverberation in common environments is often too subtle to affect human perception, it plays a critical role in adversarial audio attacks in the physical world. For instance, when attackers replay adversarial audio examples to compromise speech recognition models, the typically weak perturbations can result in attack failure due to reverberation distortion. To address this challenge, attackers must measure the RIR at the precise locations of the adversarial audio transmitter and victim microphone, then apply the RIR filter to the original speech to ensure the target device receives the intended perturbation [16, 41].

3 Threat Model

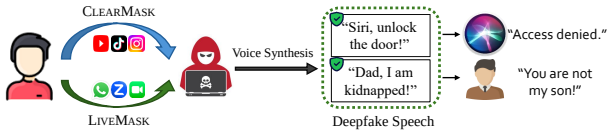
Current state-of-the-art voice cloning models are capable of synthesizing high-quality speech using just a few seconds speech sample as a reference, making it particularly challenging to effectively defend against such attacks. In this section, we define the threat model of voice deepfake attacks, elaborating on the attacker's knowledge and methods for acquiring reference voice samples, as well as the defender's capabilities of mitigating potential voice deepfake threats.

3.1 Adversary Capability

Approaches: As illustrated in Fig. 2(a), people commonly expose their voices in two scenarios. On one hand, *offline speech*, such as videos shared on public social media platforms like Instagram, TikTok, and YouTube, can be exploited by attackers who download these videos and use the speech to impersonate the user's identity. On the other hand, *online speech* from real-time applications, such as online meetings and phone calls, also puts the user's voice at risk.



(a) Voice deepfake attacks scenarios.



(b) With ClearMask and LiveMask protection, the synthesized speech could no longer deceive ASV or human ears.

Figure 2: Compared with raw speech, ClearMask protected speech is robust against voice deepfake attacks.

Protecting speech in online scenarios is particularly challenging, as users lack sufficient time to optimize the masking of these speech samples. Once an attacker successfully clones a victim's voice, the synthesized speech can be used to control voice assistants or deceive human listeners, leading to potentially severe consequences such as financial losses or physical threats [47].

Knowledge: Adversaries are capable of synthesizing deepfake speech by leveraging readily available open-source speech synthesis models or commercial APIs. Reference speech can be obtained from public social platforms or recorded from live-streaming audio. Moreover, attackers may possess private information about victims, such as their names and contact details, enabling them to execute more sophisticated attacks, including spam phone calls. Additionally, if attackers become aware that speech samples have been protected, they may attempt to bypass the protective measures and recover the raw speech. Alternatively, they could synthesize the victim's voice using various models and select the one with highest performance.

3.2 Defender Capability

Approaches: Fig. 2(b) illustrates the application scenarios of ClearMask. For offline speech, defenders can utilize ClearMask, which leverages noise-free sound effects to safeguard speech prior to uploading it onto public social media platforms. In online scenarios, defenders can activate LiveMask, the real-time mode of ClearMask to enable fast protection. With ClearMask protection, neither ASV models nor human listeners can identify the synthetic voice as belonging to the victim, thereby rendering voice deepfake attacks ineffective.

Knowledge: In this work, we consider ClearMask a *black-box* defense, meaning that defenders have no prior knowledge about the adversaries. Specifically, the model being used, including its architecture, weights, and training data, remains unknown. Furthermore, they are prohibited from querying black-box models. During the defense stage, defenders employing ClearMask can utilize public speech and RIR datasets to enhance protection performance. Moreover, they have unrestricted access to open-source voice synthesis models and their well-trained checkpoints, allowing them to build a surrogate model for the transferable defense.

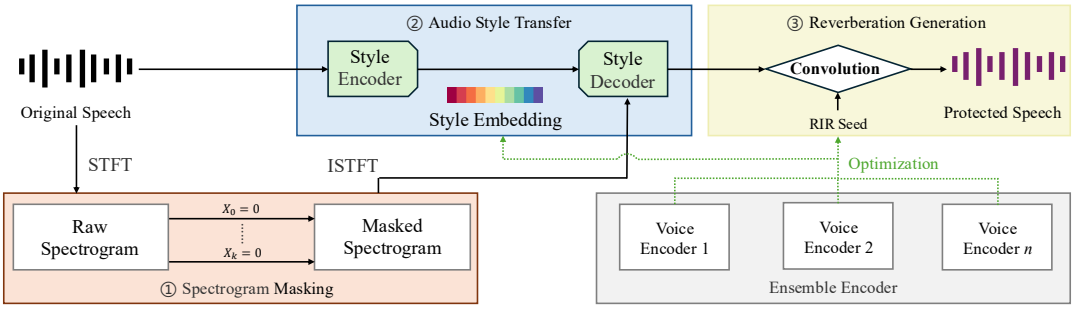


Figure 3: System overview of ClearMask. It first applies spectrogram masking to modify the mel-spectrogram. Next, an ensemble voice encoder is leveraged to optimize audio style transfer and reverberation generation stages, improving the defense effectiveness and transferability.

4 Methodology

To address the technical challenges in existing defense approaches, including transferability, audio quality, and real-time feasibility, we design ClearMask, a novel speech protection method to prevent malicious voice synthesis. The framework of ClearMask is illustrated in Fig. 3. Instead of adding noise to the original speech, ClearMask implements three techniques to mask the real voice features: spectrogram masking, audio style transfer, and reverberation generation. These noise-free methods allow us to reduce audio distortion and maintain naturalness for human perception.

4.1 Spectrogram Masking

The high sampling rate of digital audio complicates direct time-domain audio signal processing. As a result, speech synthesis models utilize spectrograms as the input. Typically, audio waveforms are converted into spectrograms using Short-Time Fourier Transform (STFT), which are then mapped to the mel scale—a perceptual scale that aligns with human auditory sensitivity by emphasizing lower frequencies. During speech generation, the decoder outputs a mel spectrogram, which is converted into an audible waveform by vocoders such as HiFiGAN [30].

Therefore, we propose masking the input mel-spectrogram of voice synthesis models to influence the output of the voice encoder E_v. Existing work [32] has demonstrated that masking certain frequencies in the spectrogram can mislead the voice feature extraction process in speech synthesis, resulting in degraded synthetic speech while preserving overall audio quality. However, this straightforward approach is not consistently effective across diverse speech samples. Unlike methods that mask wide frequency bands, ClearMask sets only a few selected frequencies to zero to minimize audio quality degradation. Furthermore, to reduce computational complexity, we first identify the frequencies with substantial power ($\geq \tau_p$), as audio spectrograms are inherently sparse matrices. Next, we apply a greedy algorithm to select the frequencies to be masked, as detailed in Algorithm 1.

The greedy frequency selection method first ranks the frequencies based on the mel-spectrogram loss $\|\text{Mel}(\mathbf{X}) - \text{Mel}(\mathbf{X}_{temp})\|_2$, where $\mathbf{X}_{temp}$ represents the spectrogram after removing a specific frequency component X_i. Next, we filter out only k frequencies with the strongest impact on the input mel-spectrogram, striking a balance between audio quality and protection performance. Notably, since the mel-spectrogram deviation serves as the loss function and does not depend on gradient back-propagation from any specific voice encoder model, this approach achieves transferable performance across diverse models.

Algorithm 1 Greedy Frequency Selection

Input: Spectrogram $\mathbf{X} = [X_0, X_1, \ldots, X_n]^T$
for $X_i \in \mathbf{X}$ **do**
  if $\|X_i\|_2 \geq \tau_p$ **then**
    $\mathbf{X}_{\text{temp}} \leftarrow \mathbf{X}$ **with** X_i **set to** 0
    $\Delta_i \leftarrow \left\|\text{Mel}(\mathbf{X}) - \text{Mel}(\mathbf{X}_{\text{temp}})\right\|_2$
  end if
end for
$\mathcal{S} \leftarrow$ indices of the top k values of Δ_i
$\mathbf{X}' \leftarrow \mathbf{X}$ **with** $X_{i \in \mathcal{S}}$ **set to** 0
return $\mathbf{X}'$

4.2 Audio Style Transfer

While masking certain frequencies can protect voice features without noticeably degrading audio quality, voice patterns can still be extracted. Therefore, additional methods are necessary to improve the effectiveness of protection. Audio style transfer, which modifies sound texture without introducing noise, is another key technique that can be utilized to spoof voice synthesis models. However, optimizing the audio style of input speech samples to achieve effective defense is challenging. Directly using another speaker's speech as $\boldsymbol{y}_t$ to process the original speech is not feasible, as explained in Appendix A. In this section, we will illustrate the audio style optimization strategy for ClearMask protection.

4.2.1 Ensemble Encoders. Existing defense methods that employ adversarial examples to attack voice synthesis models usually rely on gradient back-propagation. However, due to overfitting, the transferability of defenses is limited when the gradient from a single encoder is used to optimize the adversarial example. If the defense assumes only white-box scenarios, its effectiveness cannot be guaranteed when attackers employ alternative models to synthesize speech. Therefore, it is crucial to ensure that protected speech examples remain effective across models with varying architectures.

Fortunately, although different voice encoders are employed, various voice synthesis models share the common goal of extracting the unique vocal patterns of a single speaker from speech samples

with diverse content. Regardless of differences in model architectures, all voice encoder modules extract the same voice features, even though they are represented differently. Based on this prior knowledge, it is possible to design a transferable protection method that remains effective across different voice synthesis models. However, variations in training data and model architectures inevitably result in different gradient values. To achieve optimal performance across diverse models, ClearMask incorporates encoders with varying architectures and optimizes the loss function as follows:

$$\mathcal{L}_{spk}(x_t^{\mathcal{M}}, x_t) = \sum_{i=0}^{n} \lambda_i \| E_v^i(x_t^{\mathcal{M}}) - E_v^i(x_t) \|_2, \tag{4}$$

where E_v^i is the i-th encoder, and $x_t^{\mathcal{M}}$ is the masked speech sample generated by ClearMask. Additionally, as the dimensions of the voice embedding vectors are different, we use λ_i to adjust the weight of each encoder. This loss function based on the ensemble encoder can significantly enhance protection transferability since the optimization is guided by gradients from a diverse collection of voice encoders.

4.2.2 Style Optimization. By leveraging the surrogate ensemble encoder, we can optimize the target audio style to modify the input speech and compromise different voice synthesis models. In ClearMask, we apply DeepAFx-ST [42] as the audio style transfer tool. However, the challenge in the optimization process is that audio style transfer functions serve as a black box for voice encoder models, meaning that we cannot directly derive gradients to optimize the style embeddings. To address the challenge, we need to query the surrogate model and find the optimal target audio style for the input speech.

Moreover, in audio style transfer process, the style embedding vector $\mathbf{V}$ is normalized to have a fixed l_2 norm. Therefore, we can only flip each dimension without changing the value. We set the input of audio style transfer to be the unprocessed speech x_{in} and its style embedding vector $V = E_s(x_{in})$, along with the filtered speech $x_{in}^{\mathcal{M}}$. We use a sensitivity score to measure the efficiency of flipping each dimension in V:

$$Sen(x_{out}, x_{in}^{\mathcal{M}}, x_{in}) = \frac{\mathcal{L}_{spk}(x_{out}, x_{in}) - \mathcal{L}_{spk}(x_{in}^{\mathcal{M}}, x_{in})}{\|\mathbf{X}_{in} - \mathbf{X}_{out}\|_2}, \tag{5}$$

where $\mathbf{X}_{in}$ and $\mathbf{X}_{out}$ are spectrograms of x_{in} and x_{out}. Algorithm 2 presents the audio style optimization process.

First, we attempt to flip each dimension in V and record the sensitivity, which is the ratio of the extra voice embedding loss to the audio quality loss caused by style transfer. Next, we flip these dimensions according to the sensitivity score from high to low. Meanwhile, we set an audio distortion threshold τ. Once the audio quality loss is beyond this threshold, we will quit the loop to ensure the audio distortion is acceptable with $x_{out} = S(x_{in}^{\mathcal{M}}, V_{out})$. Although sounds similar, x_{out} causes completely different voice embedding vector from the raw speech x_{in}.

Algorithm 2 Style Embedding Optimization

Input: style vector $V = [v_0, v_1, \ldots, v_n]$, x_{in} and $x_{in}^{\mathcal{M}}$
for each $v_i \in \mathbf{V}$ **do**
  $V_{temp} \leftarrow V$ with $v_i = -v_i$
  $x_{out} = S(x_{in}^{\mathcal{M}}, V_{temp})$
  $Score[i] = Sen(x_{out}, x_{in}^{\mathcal{M}}, x_{in})$
end for
Sort $Score[i]$ in descending order
for each i in sorted $Score[i]$ **do**
  while $\|X_{in} - X_{out}\|_2 < \tau$ **do**
    $\mathbf{V}_{out} \leftarrow \mathbf{V}$ with $v_i = -v_i$
  end while
end for

4.3 Reverberation Generation

4.3.1 RIR Selection. In the final step, we enhance the protection effectiveness by adding additional reverberation to the protected speech. As mentioned in Section 2.3, reverberation is generated by convolving a reversed RIR with the raw audio signal. The selection of an appropriate RIR is guided by two main factors. On one hand, RIR signals are typically characterized as damped oscillatory signals with a rapid initial decay. To maintain the naturalness of the processed audio, the reverberation follows this characteristic. On the other hand, while reverberation is ubiquitous in physical environments, excessive reverberation can still degrade audio quality. Therefore, we must impose constraints on RIR seeds for reverberation generation.

First, the length of the RIR seed determines the perceptibility of the reverberation. Typically, when the sound reflection delay is shorter than 30 ms, the reflected sound is mixed with the original sound [23] so that it is imperceptible. When the delay is longer, human ears identify the reflected sound as an echo, causing a significant difference in human perception. Therefore, ClearMask constrains all RIR seeds length within 30 ms to guarantee audio clarity. Second, to ensure naturalness, we collect RIR seeds from reverberation audios recorded in various real-world environments [43]. We normalize all RIR samples and initialize the RIR seeds by clipping only the initial part of the RIR signal with relatively strong power. Next, we attempt to select the RIR with the strongest protection performance, h^*, to maximize the loss of voice embedding vector, as shown in Eq. (4):

$$h^* = arg \max_{h \in S^{RIR}} [\mathcal{L}_{spk}(x_{out} * h, x_{in}) - \lambda_l \cdot len(h)], \tag{6}$$

where h represents the reversed RIR signals in the dataset S^{RIR}. Meanwhile, we attempt to shorten the RIR length by incorporating a penalty term, $-\lambda_l \cdot len(h)$, to punish excessively long reverberation in the protected speech audio.

4.3.2 Reverberation Optimization. Natural reverberation alone may not always provide sufficient protection against malicious voice synthesis. Therefore, the natural RIR seeds require further optimization to achieve the highest protection performance. At this stage, the loss applied for optimization is different. The previous goal of ClearMask optimization stages is to maximize the loss between the real voice embedding and the protected voice embedding. However, continuously increasing the loss value does not always lead to better protection. When the loss exceeds a certain threshold, the synthesized speech becomes noisy rather than merely altering the vocal features. This effect may alert adversaries to the

presence of protective measures. Our solution is to select a "target speaker" with a completely different voice from the protected speaker to effectively mislead the voice encoder. AntiFake [52] applies a human-in-the-loop method, which asks human operators to select speech samples that are most dissimilar to the protected voice to constrain the voice embedding loss. However, this method is labor-intensive, particularly in scenarios requiring large-scale protection. Our approach addresses this limitation by employing a speaker recognition model to extract voice features and automatically select samples that are most dissimilar to the protected voice. After it, the maximization problem is converted to a minimization problem. The optimization can then be formulated as:

$$\min_{\delta} [\mathcal{L}_{spk}(x^{tgt}, x_{out} * (h^* + \delta)) - \lambda \cdot \mathcal{L}_{spk}(x_{in}, x_{out} * (h^* + \delta))] \quad \textbf{s.t.} \quad \|\delta\|_\infty < \epsilon, \tag{7}$$

where x^{tgt} represents the speech sample from the target speaker, and λ adjusts the weight of loss values. Additionally, by introducing ϵ, we can constrain the RIR amplitude and mitigate the interference caused by reverberation.

In addition, we optimize the projected gradient descent (PGD) method to enhance protection transferability. We notice that, at the beginning of PGD, the adversarial examples show good transferability across different models. However, as the PGD optimization continues, transferability tends to degrade. After running PGD for multiple iterations, for $\mathcal{L}_{spk}$ in Eq. (4), the loss values of some encoders continue to decrease, while the loss values from other voice encoders remain unchanged or even increase. To avoid this transferability loss, we implement a new strategy: for each individual voice encoder in the ensemble, if its voice embedding loss does not decrease over k_c consecutive iterations, we stop the iteration. This prevents the optimization process from overly focusing on a single model, which reduces overall protection transferability. Furthermore, because our approach incorporates three different stages and a surrogate model composed of multiple encoders, it effectively enhances protection robustness. No voice synthesis model can withstand all of the protective methods or encoders.

4.4 LiveMask Design

While we can effectively protect offline speech , online streaming speech, such as online meetings or voice messages, is still vulnerable. Compared to offline protection, online protection requires minimal latency, making step-by-step optimizations infeasible. To address this challenge, we propose a fast protection model of ClearMask, named LiveMask, to extend its feasibility for online applications. In online protection, we skip audio style transfer because it requires the entire speech sample before processing. In contrast, spectrogram masking and reverberation can be rapidly applied to real-time audio signals, enabling fast protection.

Prior to the optimization, we prepare a dataset $\mathcal{D} = \{\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$, which contains 150 seconds speech samples from a single user, covering common English phonemes. In the first step, we optimize a general frequency mask $\mathcal{M}_g$ using the following objective:

$$arg\max_{\mathcal{M}_g} \sum_{i=0}^{n} \|Mel[\mathcal{M}_g(\boldsymbol{x}_i)] - Mel(\boldsymbol{x}_i)\|_2, \tag{8}$$

where $\mathcal{M}_g$ is a filter that masks k fixed frequencies $f_0, f_1, ..., f_{k-1}$ to maximize the mel-spectrogram loss across all samples in $\mathcal{D}$. This mechanism pre-configures the filtered frequencies for the speaker. In this way, when the microphone captures the streaming speech, the frequency filter can be applied in real-time without introducing additional latency.

Moreover, we design an optimization process for a universal RIR seed h_g to generate reverberation in the speech. Similar to the RIR optimization in ClearMask, this universal RIR seed h_g is designed to minimize $\mathcal{L}_{spk}$ across all samples in $\mathcal{D}$. The optimization process is formulated as:

$$arg\min_{\delta_g} \sum_{i=0}^{n} [\mathcal{L}_{spk}(x^{tgt}, x_i' * (h + \delta_g)) - \mathcal{L}_{spk}(x_i, x_i' * (h + \delta_g))] \quad \textbf{s.t.} \quad \|\delta_g\|_\infty < \epsilon, \text{ and } h_g = h + \delta_g, \tag{9}$$

where $\boldsymbol{x}_i'$ represents the speech samples after universal frequency masking $\mathcal{M}_g$. It is notable that finding the optimal solution for LiveMask is more critical, as the universal RIR seed should indiscriminately protect ambient speech containing different contents from the given speaker. To achieve this, we decrease the learning rate and increase the number of iterations to ensure the model finds a solution with better generalized performance. Once the universal RIR seed is determined, the reverberation can be immediately applied to streaming speech signals via convolution. Finally, the latency corresponds to the length of the RIR, which is typically tens of milliseconds. Also, as we bypass the optimization process for individual speech samples, the masked universal filtered frequencies k and the δ_g constraint ϵ in LiveMask are less strict compared to the offline mode of ClearMask.

Table 2: Voice synthesis models in ClearMask.

	Voice Synthesis Models	Encoder Architecture	Embedding Dimension
Surrogate models	AdaIN-VC [19]	VAE	128
	AutoVC [38]	VAE	256
	SV2TTS [45]	LSTM	256
Test models	YourTTS [13]	ResNet	512
	DiffVC [36]	VAE	256
	AGAIN-VC [18]	U-Net	N/A
	ElevenLabs [3]	N/A	N/A
	Play.ht [5]	N/A	N/A

5 Evaluation

5.1 Experiment Setup

5.1.1 Speech Datasets. We use two datasets for evaluation: one is VCTK-Corpus [44], an English dataset that contains 107 speakers with different accents. Meanwhile, since most voice synthesis models are originally trained on this dataset, which may introduce bias into the results, we also use LibriSpeech [34], another English speech dataset with longer speech samples. We randomly select 50

speakers from each dataset (100 speakers in total) and test the 20 longest speech samples from each speaker.

5.1.2 Speech Contents. We use ChatGPT 4.0 [7] to generate 20 textual inputs that could potentially be used in voice deepfake attacks for TTS-based synthesis. For VC-based synthesis methods, we use Google Text-to-Speech [8] to generate speech from these text inputs as the source speech. Some textual input samples are included in Appendix B.

5.1.3 Voice Synthesis Models. We employ multiple voice synthesis models for both the training and testing stages. For surrogate models, we utilize AdaIN-VC [19], AutoVC [38], and SV2TTS [45], which are based on Variational Autoencoder (VAE) and Long Short-Term Memory (LSTM). We use 3 different open-source voice synthesis models for testing. YourTTS [13] is a TTS model with a ResNet-based voice encoder. DiffVC [36] is a voice conversion model using diffusion model to generate speech mel-spectrogram. AGAIN-VC [18] is a voice conversion model using U-Net architecture to extract the content along with voice embedding vectors. The details of these models are summarized in Table 2, where N/A indicates parameters that are either unknown or not applicable. Although the basic architectures of these models are similar, their detailed parameters, such as the number of layers and weights, differ significantly, requiring ClearMask protection to achieve effective transferability. Additionally, we test ClearMask performance on two commercial voice synthesis APIs: ElevenLabs [3] and Play.ht [5]. These commercial APIs operate as pure black boxes, meaning we have no knowledge about their model architectures or weights.

5.1.4 Hyperparameters. We resample all speech audio sampling rate to 48 kHz in the pre-processing stage. In spectrogram masking, we use STFT to transfer the audio waveform to a spectrogram with 1025 frequency bins mask k=12 frequencies.

5.1.5 Evaluation Metrics. We use three different metrics to comprehensively evaluate ClearMask performance on ASV:
Similarity Score: SpeechBrain [40] is an open-source speech toolkit built on PyTorch [35]. Its speaker verification function is built on the state-of-the-art ECAPA-TDNN [20] model. The similarity score is calculated using the cosine similarity between the reference voice embedding and the inference voice embedding vectors.
ECAPA-TDNN Rejection Rate (ETRR): The ECAPA-TDNN model provides a speaker verification decision based on the similarity score. If the similarity score falls below a certain threshold (default = 0.25), the voice is rejected by ASV.
Soniox Rejection Rate (SRR): We also evaluate ClearMask performance using Soniox [9], a speaker identification service API. We first enroll the original speech samples as the reference and then upload the synthesized speech. If the synthesized voice is identified as an "unknown speaker," we consider ClearMask to have successfully prevented the voice deepfake attack.

5.2 ClearMask Effectiveness on Open-source Voice Synthesis Models

First, we evaluate the performance of ClearMask on unseen voice synthesis models. To test the synthesis capabilities of the models, we use unprotected speech samples as reference inputs for open-source voice synthesis models and evaluate the similarity between the synthesized and original voices using the ECAPA-TDNN model. The results are listed in the first row of Table 3. Although the synthesis performance varies across different models, most synthetic voice samples are successfully verified by the ECAPA-TDNN speaker verification system across all models. This demonstrates that sharing unprotected speech online can easily compromise ASV systems.

Next, we conduct an ablation study by comparing each step in ClearMask individually and in different combinations. When we apply only the spectrogram masking method, the similarity between the synthesized voice samples and the unprotected synthesized voices significantly decreases across all models, indicating good transferability of this approach. However, despite the reduction in similarity, many synthesized voice samples can still pass speaker verification as voice features still remain in the residual frequencies. Moreover, further increasing the number of masked frequencies would result in a substantial decline in audio quality. Therefore, simply relying on this method is insufficient to comprehensively protect human speech against voice synthesis. In addition, we evaluate the protective performance of the audio style transfer and reverberation generation methods. Compared to spectrogram masking, these two methods are optimized based on the loss provided by the surrogate ensemble encoders. In our experiments, both methods demonstrate high protective effectiveness. Despite significant differences in structure and parameters across different models, the aggregate encoder approach we use has a noticeable impact on all models. After applying audio style transfer, most synthesized voice samples can no longer bypass speaker verification. Similarly, the reverberation generation method successfully protects over 80% of the speech samples. However, when used individually, these methods can still be compromised by attackers by repeated attempts.

For comparison, we run experiments with different combinations of these three methods. As shown in Table 3, the protective effectiveness for unseen models significantly increases when multiple methods are applied. When two of the methods are used in combination, at least 94% of the voice samples are successfully protected. When all three methods are combined, nearly all samples are successfully protected across all potential attack models. Considering that the ECAPA-TDNN model has an error rate of approximately 1%, ClearMask remains effective even if the rejection rate is not always 100%[1].

Meanwhile, we evaluate the effectiveness of ClearMask on Soniox, a commercial speaker recognition API. Generally, commercial ASV models have more complex architectures, leading to better robustness than open-source models. Since commercial APIs will be used in real-world application scenarios to process diverse speech data, the ASV model must ensure the lowest possible false acceptance rate (FAR), which means that it has a higher threshold to reject deepfake or similar voice samples. Even generated from unprotected samples, a lot of synthetic voices fail to pass the commercial ASV models. When we apply different protection combinations, most methods successfully achieve a 100% protection success rate. Therefore, ClearMask demonstrates strong effectiveness in preventing all synthetic speech samples from passing commercial ASV models.

[1]Audio demos are available at: https://clear-mask.github.io/.

Table 3: ClearMask performance on unseen open-source voice synthesis models.

	Defenses	YourTTS			DiffVC			AGAIN-VC		
		Score ↓	ETRR ↑	SRR ↑	Score ↓	ETRR ↑	SRR ↑	Score ↓	ETRR ↑	SRR ↑
Ablation Study	N/A	0.577±0.19	4.2%	26.4%	0.532±0.22	6.6%	29.0%	0.366±0.16	11.3%	44.0%
	①	0.202±0.09	75.6%	91.2%	0.209±0.09	71.3%	90.0%	0.177±0.07	84.0%	96.5%
	②	0.226±0.12	68.0%	90.5%	0.211±0.10	70.6%	92.9%	0.189±0.11	82.2%	97.0%
	③	0.187±0.08	83.5%	94.5%	0.188±0.11	88.4%	96.9%	0.158±0.13	95.0%	99.1%
	①+②	0.176±0.09	94.2%	100%	0.164±0.08	95.3%	100%	0.120±0.06	100%	100%
	①+③	0.159±0.07	98.4%	100%	0.147±0.06	98.0%	100%	0.109±0.06	100%	100%
	②+③	0.163±0.07	96.5%	100%	0.152±0.07	98.8%	100%	0.114±0.05	100%	100%
	①+②+③	**0.125±0.05**	**99.8%**	**100%**	0.112±0.04	99.9%	**100%**	0.091±0.04	**100%**	**100%**
Existing Defenses	Attack-VC [24]	0.233±0.18	64.2%	78.0%	0.227±0.17	76.3%	84.1%	0.194±0.12	82.5%	93.9%
	SampleMask [32]	0.286±0.22	46.6%	58.4%	0.266±0.20	53.2%	69.0%	0.210±0.16	70.5%	86.8%
	AntiFake [52]	0.138±0.06	**99.8%**	**100%**	**0.107±0.04**	**100%**	**100%**	**0.085±0.04**	**100%**	**100%**

① Spectrogram Masking ② Audio Style Transfer ③ Reverberation Generation

Table 4: LiveMask performance on unseen open-source voice synthesis models.

Defenses	YourTTS			DiffVC			AGAIN-VC		
	Score ↓	ETRR ↑	SRR ↑	Score ↓	ETRR ↑	SRR ↑	Score ↓	ETRR ↑	SRR ↑
①	0.211±0.16	70.2%	88.5%	0.221±0.18	62.6%	80.4%	0.188±0.13	79.4%	91.7%
②	0.182±0.12	83.5%	94.4%	0.169± 0.12	89.4%	97.5%	0.155±0.10	93.6%	100%
①+②	**0.154±0.08**	**99.7%**	**100%**	**0.145±0.07**	**99.9%**	**100%**	**0.126±0.05**	**100%**	**100%**
VSMask [46]	0.226±0.17	70.8%	84.6%	0.239±0.20	66.0%	79.5%	0.181±0.14	85.8%	95.2%

① Universal Spectrogram Masking ② Universal Reverberation Generation

In addition, we compare the performance of ClearMask with other baseline offline defense methods, including Attack-VC, SampleMask, and AntiFake. For Attack-VC and SampleMask, we adopt the white-box defense setup. We choose the samples with the best transferability performance across multiple target voice synthesis models. As shown in Table 3, when Attack-VC and SampleMask are applied to defend against unseen voice synthesis models, they fail to achieve effective results due to their inability to handle varying model structures and parameters. After applying these defenses, the synthesized voices still exhibit high similarity to the original voices, failing to meet the transferability requirements for defending against deepfake voices. In comparison, AntiFake leverages an ensemble encoder approach to improve transferability, and our experiments confirm its effectiveness. Similar to ClearMask, AntiFake achieves nearly 100% defense success rates in black-box scenarios. However, AntiFake suffers from significant degradation in the quality of the protected audio, which we will compare in detail in Section 5.6.

5.3 LiveMask Effectiveness Evaluation

Next, we apply LiveMask protected speech to the same open-source voice synthesis models. As introduced in Section 4.4, LiveMask does not involve optimization for individual speech samples. Therefore, we modify the constraint hyperparameters used in ClearMask. In the universal spectrogram masking stage, we mask $k = 16$ fixed frequencies. In the reverberation generation stage, the length of h_g is fixed to 30 ms.

Similar to offline ClearMask evaluation, we assess the performance of each stage and LiveMask with both stages in the ablation study. The experimental results are shown in Table 4. In the spectrogram masking step, the universal protection method only masks fixed frequencies. However, removing more frequencies is not feasible given the audio quality requirements. Additionally, due to the diversity of human languages, this fixed protection strategy cannot guarantee effectiveness for all speech content. Consequently, we observe that while universal spectrogram masking reduces the similarity between synthesized and original speech compared to unprotected speech, as shown in Table 3, the average similarity still remains relatively high, and the standard deviation is noticeably greater than ClearMask. Such result indicate that for some speech samples, only masking fixed frequencies is not sufficiently effective. Similarly, the universal reverberation generator can also reduce the similarity between synthesized voices and the original voice, but it fails to provide complete protection. In comparison, when we combine the two stages, the protection demonstrates better transferability and effectiveness across all models. At least 99.7% of cloned voices cannot bypass the ECAPA-TDNN ASV model, and none of the samples can spoof the Soniox ASV model. Therefore, LiveMask achieves effective protection across various unseen voice synthesis models. We further assess the computational requirements of LiveMask. Specifically, when applying reverberation to a 48 kHz audio signal, the computational cost is 70 million floating-point operations per seconds (FLOPs), with a memory footprint of 80 MB. These results demonstrate that LiveMask is capable of running on lightweight devices, such as the Raspberry Pi 4, which features at least 1 GB of random-access memory (RAM) and supports 3 billions FLOPs. Moreover, the computational efficiency of LiveMask can be further optimized for IoT devices by reducing the audio sampling rate, thereby lowering both processing and memory demands.

Moreover, we compare LiveMask with the existing online protection method, VSMask. We find that VSMask is ineffective against unseen voice synthesis models. This is because its predictive model is trained in a white-box setup and heavily depends on the weights of a specific voice synthesis model. As a result, the generated perturbations fail to maintain their effectiveness across different models.

Table 5: ClearMask and LiveMask performance on commercial voice synthesis platforms.

Method	ElevenLabs		Play.ht	
	ETRR ↑	SRR ↑	ETRR ↑	SRR ↑
N/A	1.0%	19.3%	0.0%	12.2%
LiveMask	98.8%	100%	97.4%	100%
ClearMask	**99.2%**	**100%**	**98.3%**	**100%**

5.4 ClearMask and LiveMask Performance on Commercial Platforms

Next, we evaluate the performance of ClearMask and LiveMask on commercial voice cloning platforms. Similar to ASV models, commercial voice cloning services typically use larger-scale models and more extensive training data than open-source voice synthesis models, resulting in greater robustness. Moreover, these models are entirely black-box, meaning we can only counter them by enhancing the transferability of our protection methods.

Compared to open-source speech synthesis models, commercial models require longer speech samples to improve synthetic speech quality. Therefore, we concatenate the protected samples into longer voice samples. We compare the voice rejection rates of the ECAPA-TDNN and Soniox ASV models before and after applying ClearMask and LiveMask protection. The experimental results are presented in Table 5. According to the results, commercial voice cloning platforms achieve lower voice rejection rates than open-source voice synthesis models. Almost all synthetic speech samples generated from unprotected speech can successfully spoof the ECAPA-TDNN model, and over 80% can even bypass the Soniox speaker recognition API, revealing a severe threat to user privacy. Additionally, because these commercial platforms do not require any local resources, such as GPUs or development environments, they present a greater threat than open-source models. With just 10 seconds of unprotected clear speech, these commercial models can generate a large number of high-quality and highly similar synthetic speech samples.

Next, we upload voice samples protected by ClearMask to the API and test the similarity between the generated voices and the original voices. Despite having no knowledge of the internal workings of the API models, ClearMask still demonstrates high transferability. Over 98% of the voice samples generated from ClearMask protected samples cannot pass the ECAPA-TDNN based ASV model, and 100% are rejected by the commercial speaker recognition API. Additionally, LiveMask can also successfully protect over 97% of samples in real-time. As a result, both ClearMask and LiveMask demonstrate strong protection effectiveness on black-box voice synthesis API platforms.

5.5 Human Perception Evaluation

To test the performance of ClearMask for prevent synthetic speech from spoofing human perception, we use perceptual speaker dissimilarity (PSD) to measure it. The PSD score ranges from 1 to 5, where 1 indicates that the two voices are from the same speaker, and 5 indicates that the two voices are distinct. To save human effort, we select the best-performing open-source and commercial voice cloning models, YourTTS and Play.ht, as the adversarial models to verify the effectiveness of ClearMask and LiveMask. We collected over 550 responses from 28 listeners (17 males and 11 females, aged 20 to 35) with normal hearing ability, who are asked to rate the PSD between the two voice samples.

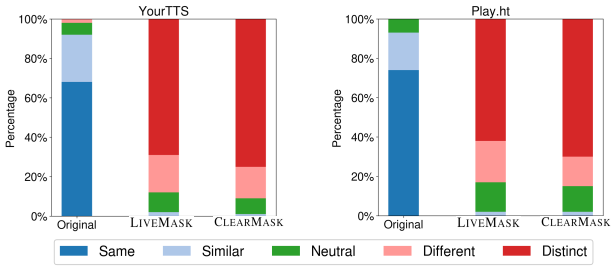


Figure 4: PSD comparison of synthetic voice generated from unprotected and protected speech samples.

We present the PSD results in Fig. 4. When the unprotected voice samples are exposed to attackers, the synthetic speech from advanced voice synthesis models can easily deceive human listeners. Over 90% of responses indicate that the synthetic voice is the same as or similar to the reference voice. These synthetic speech samples are highly likely to fool listeners into misjudging the speaker's identity, potentially leading to financial losses or security threats. In comparison, with ClearMask and LiveMask protection, the synthetic voice sounds distinct from the real voice. Less than 2% of the samples are identified as "similar to" the real voice. None of the samples are considered to be from the same speaker as the reference voice. When attackers use protected speech to synthesize deepfake voice, the resulting speech not only exhibits different characteristics from the original voice but also sounds hoarse and unstable. As a result, the synthetic speech can be easily recognized by human listeners as not being from the target speaker, resulting in a failed attack.

5.6 ClearMask Audio Quality Comparison with AntiFake

One of the main innovations of ClearMask and LiveMask is that the strategies we apply, including spectrogram masking, audio style transfer, and reverberation generation, do not introduce ambient noise to the original clear speech. This allows ClearMask and LiveMask to achieve better audio quality and naturalness compared to other existing voice deepfake defenses based on perturbation optimization. In this section, we use different evaluation metrics to measure the audio quality of protected speech samples from various protection methods.

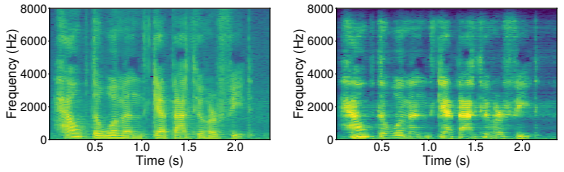
Mean Opinion Score (MOS): A measurement of audio quality based on subjective listener ratings. We use NISQA, a DNN-based method, to estimate speech quality and naturalness on a scale from 1 to 5. Typically, when the MOS is higher than 3, the speech is considered clear for human perception.

Short-Time Objective Intelligibility (STOI): STOI measures the intelligibility of the processed input signal by comparing it with the clean reference signal. The STOI score ranges from 0 to 1, representing speech that is absolutely unintelligible to highly intelligible.

Perceptual Evaluation of Audio Quality (PEAQ): PEAQ is a

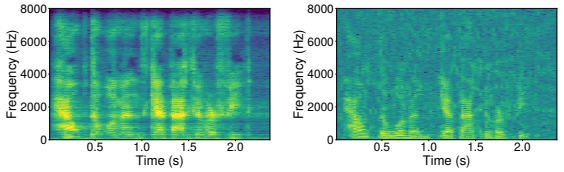
Table 6: ClearMask and LiveMask protected audio quality comparison with AntiFake.

	MOS ↑	STOI ↑	PEAQ ↑
AntiFake [52]	2.82±0.66	0.34±0.11	2.41±0.83
LiveMask	3.01±0.80	0.59±0.12	3.77±0.77
ClearMask	**3.12±0.72**	**0.65±0.09**	**4.26±0.51**



(a) The original speech.

(b) The speech protected by ClearMask.



(c) The speech protected by LiveMask.

(d) The speech protected by AntiFake.

Figure 5: Spectrograms of raw speech and protected speech samples from ClearMask, LiveMask and AntiFake.

subjective audio quality evaluation based on real human perception. Listeners evaluate their perceived audio quality and provide an opinion score on a scale from 1 (very noisy) to 5 (imperceptible noise).

Considering that white-box defense speech samples may have better audio quality, this trade-off compromises their ability to achieve transferable defense, making them less practical for use. In the evaluation, we only compare AntiFake [52] with ClearMask and LiveMask. The comparison results are listed in Table 6. For MOS, ClearMask is slightly better than LiveMask and the AntiFake methods. Although ClearMask and LiveMask do not apply noise injection, the MOS measurement is sensitive to audio "coloration," which refers to artificial patterns. Therefore, additional reverberation can also degrade the opinion score. In contrast, ClearMask and LiveMask show much better results in STOI and PEAQ. In Fig. 5, we display spectrograms from unprotected speech sample along with the protected samples generated by ClearMask, LiveMask, and AntiFake. We observe that while ClearMask inevitably causes some loss in audio quality compared to unprotected samples, the protected audio still retains most of its original features. As a result, ClearMask achieves relatively high STOI and PEAQ scores. In contrast, LiveMask introduces additional reverberation and signal loss compared to ClearMask, resulting in slightly lower scores. By comparison, in AntiFake-protected audio, noise is dispersed across the entire frequency spectrum, with high-frequency signals above 4 kHz almost entirely overwhelmed by adversarial noise. Although AntiFake attempts to mitigate the impact of noise based on human auditory sensitivity, it still severely degrades the quality and intelligibility of the sound, making it unsuitable for applications where high audio quality is required.

5.7 Adaptive Attackers

We consider two types of adaptive attackers. The first type, referred to as **R1**, has no prior knowledge and attempts to recover the original audio using conventional signal processing techniques, such as WaveGuard [26]. The second type, referred to as **R2**, is aware of ClearMask's defense mechanism and employs deconvolution methods to remove reverberation and restore the original audio.

R1 Attacker: We implement four transformations in WaveGuard: Quantization-Dequantization (Quant.), Down/Up-sampling (Resamp.), Frequency Filtering (Filter.), and Mel-spectrogram Inversion (Mel.). Specifically, we quantize the audio to 8 bits and down-sample it to a frequency of 8 kHz.

The results are presented in Fig. 6. Compared to the raw protected audio (Baseline), WaveGuard-processed samples fail to recover the voice features. None of these transformations enable synthetic speech to bypass the ASV model. This is because WaveGuard attempts to eliminate subtle perturbations by compressing the audio features. However, these approaches are ineffective against the defense mechanisms in ClearMask, as the protected audio contains no noise-based perturbations to exploit.

R2 Attacker: For more sophisticated attackers who have knowledge about ClearMask, they attempt to remove the adversarial reverberation effect in the protected speech. In this experiment, the RIR length in ClearMask is fixed at 30 ms. However, the length is an unknown hyperparameter for the adaptive attackers, so they would adopt the same ensemble encoder to optimize RIRs of varying lengths for deconvolution.

In Fig. 7, we show the impact of adaptive attackers using YourTTS for deepfake voice generation. We find that regardless of the length of the natural RIR used by adaptive attackers for deconvolution to attempt to remove reverberation, they consistently fail to generate qualified cloned voices from the processed speech samples. The reason is that both spectrogram masking and audio style transfer play critical roles in ClearMask protection. Moreover, these techniques are irreversible, meaning that even with comprehensive prior knowledge, attackers cannot mitigate their effects. Second, once the speech sample is protected with ClearMask, the voice embedding vector becomes similar to an unknown "target speaker" voice. In this way, adaptive attackers are unaware of the real voice embedding and will attempt to push the voice embedding vector far from its initial state. Thus, this attack approach will not succeed in recovering the genuine voice features.

Moreover, adversarial training can enhance the robustness of speech synthesis models against protected samples. To test its feasibility, we apply adversarial training to AdaIN-VC using adversarial samples generated via PGD with 1,000 iterations. However, this process requires 150 times more training time (over 20 days) than standard training, making it excessively costly and potentially impractical for attackers. Additionally, it further degrades the model's ability to synthesize ordinary speech [33]. Therefore, we do not discuss it in detail.

6 Discussion

ClearMask Protection Performance in Other Languages. In Section 5, we evaluated the protection performance of ClearMask and LiveMask. However, since different languages exhibit distinct phonetic characteristics, it is essential to test the effectiveness of

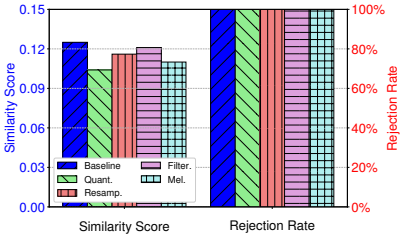


Figure 6: ClearMask performance when adaptive attackers apply WaveGuard to recover the speech.

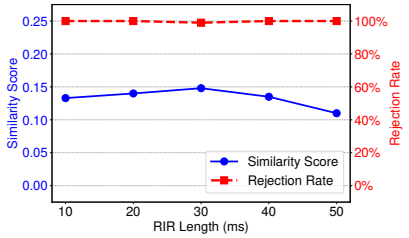


Figure 7: ClearMask performance against adaptive deconvolution-based reverberation removal.

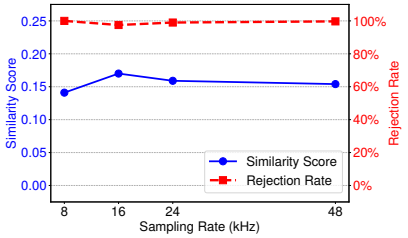


Figure 8: LiveMask performance comparison under different audio sampling rates.

Table 7: ClearMask and LiveMask protection performance in different languages.

Language	Raw Speech		ClearMask		LiveMask	
	ASS ↓	ETRR ↑	ASS ↓	ETRR ↑	ASS ↓	ETRR ↑
English	0.577	4.2%	0.125	100%	0.154	99.8%
Mandarin	0.496	7.6%	0.119	100%	0.151	99.6%
Spanish	0.520	5.5%	0.128	100%	0.159	99.0%

ClearMask and LiveMask in other languages as well. In this experiment, we test ClearMask and LiveMask in Mandarin and Spanish, two of the most widely spoken languages in the world besides English, to validate the transferability of ClearMask across multiple languages. All datasets are collected from the Mozilla Common Voice [11]. We use YourTTS to generate deepfake voices, as its foundation model is trained on multiple languages to ensure that the speech synthesis capability remains consistent across different target languages.

We use average similarity score (ASS) and ETRR to measure the protection effectiveness. The experimental results are listed in Table 7. When no protection is applied, the model's ability to synthesize Mandarin and Spanish speech is slightly weaker than English, due to differences in the amount of training data. Nevertheless, the synthesized speech still exhibits high voice similarity, with over 90% of the voice samples successfully passing ECAPA-TDNN speaker verification. In contrast, when we apply ClearMask and LiveMask protection to the reference voice samples, the similarity of the synthesized Mandarin and Spanish voice samples is significantly reduced. This demonstrates that our protection strategy is effective across different languages, even when their pronunciation and intonation are different.

LiveMask Performance under Different Sampling Rates. In real-world scenarios, the audio sampling rate for real-time voice communication is constrained by network speed and hardware limitations, including microphones. To assess the effectiveness of LiveMask under different conditions, we evaluate its performance across various sampling rates.

As illustrated in Fig. 8, we test audio sampling rates of 8 kHz, which is used in telephone communication, and higher rates including 16 kHz, 24 kHz, and 48 kHz. Next, we synthesize deepfake voices from resampled audios on YourTTS model. It is important to note that we do not design distinct reverberations for different sampling rates. Instead, we down-sample the RIR according to the audio's sampling rate. From the results, we observe that LiveMask achieves optimal defensive performance at the default 48 kHz sampling rate. When the audio is sampled at 16 kHz, the defensive performance has a slight decline but still maintains a protection success rate above 97%. At 8 kHz, due to the significantly reduced sampling rate, the effectiveness of deepfake voice synthesis is inherently diminished. Overall, these results demonstrate that ClearMask delivers robust protective performance across a range of hardware and network conditions, maintaining its effectiveness even under varying sampling rate constraints.

ClearMask Latency. We also compare the latency of ClearMask with other defense methods. When the number of iterations is set to 1000, Attack-VC [24] takes an average of 35 seconds to generate a protected sample. AntiFake [52] takes about three minutes to provide protection due to the computational cost of gradient calculations across multiple models. Additionally, its human-in-the-loop mechanism for voice similarity assessment adds further latency. In contrast, while ClearMask also computes gradients across multiple models, it optimizes an RIR signal of under 1,440 samples at 48 kHz. Other methods modify the mel spectrogram, requiring gradient calculations over thousands of dimensions, making ClearMask significantly faster than AntiFake. The average time for ClearMask reverberation generation is 40 seconds, and when combined with the computation time for spectrogram masking and audio style transfer, the total average time is 60 seconds. For LiveMask, since frequency filtering and convolution can be performed in real time, we consider its latency to be the length of the RIR signal, which is approximately 30 milliseconds.

Ethical Statement. We are deeply committed to addressing potential ethical issues associated with this research. We exclusively utilize publicly available speech datasets and open-source speech synthesis or speaker verification models, following their usage guidelines and licenses. All human listeners have provided informed consent after being informed of the study's purpose and procedures before participation, and the experiments involving human subjects have received IRB approval. Their privacy and personal information are also rigorously protected throughout the research process.

7 Related Work

7.1 Voice Synthesis and Countermeasures

Many VC and TTS models have been developed to embed a voice onto given linguistic content. Early voice conversion methods relied on parallel speech data for model training [17]. Recent VAE-based voice conversion approaches are able to learn from different speech

contents from various speakers [19]. AutoVC [38] introduces a bottleneck mechanism to enable zero-shot voice conversion. Different from VC, TTS can generate speech only according to textual input. SV2TTS [27] can embed ambient speaker's voice into the synthetic speech. Currently, multiple TTS platforms support high-quality and human-alike speech generation [3, 5].

Although voice synthesis is rapidly developing, it remains susceptible to adversarial examples, as it is fundamentally based on deep learning technology. Attack-VC [24] is the first work leveraging adversarial examples to spoof voice synthesis models, misleading them into producing unqualified voice samples. Despite its effectiveness on the target voice synthesis model, this approach requires long computation time to optimize adversarial perturbations, making it impractical for real-time speech protection. VSMask [46] attempts to address this issue by designing a predictive model that forecasts perturbations for upcoming streaming speech. While it overcomes the latency issue caused by offline optimization, the approach still depends on gradient back-propagation to optimize the predictive model, degrading its black-box performance. SampleMask [32] employs different types of frequency masks to protect speech samples while preserving their quality and achieves black-box protection by querying voice synthesis models. However, its protection is limited in transferability, as it targets only a single model. Furthermore, the protection is not foolproof, leaving some samples vulnerable to attacks if the attacker attempts multiple times. AntiFake [52] offers effective and transferable protection across various voice synthesis models. Nevertheless, since it injects optimized noise into the raw speech sample, it struggles to protect samples in real-time while maintaining their naturalness and clarity. In contrast, ClearMask overcomes the limitations of existing methods with a noise-free and universal protection mechanism, providing a robust solution for both online and offline applications.

7.2 Audio Style Transfer

Audio styles, characterized by elements such as timbre, spatialization, and loudness, present complex variables that are challenging to manually manipulate. Recent developments in automatic audio generation models enable intelligent audio style transfer with high performance and efficiency. Drawing inspiration from image style transfer [21], Grinstein et al. [22] propose an innovative audio style transfer framework utilizing convolutional neural networks. In addition, the approach of differentiable signal processing marks a significant enhancement in this field [39], optimizing the audio style transfer framework through backpropagation and effectively addressing the issue of weak correlation between audio and parameters. DeepAFx-ST [42] applies differentiable signal processing along with self-supervised training, achieving audio style transfer without reliance on labeled data. Meanwhile, SpeechSplit [37] provides a speech decomposition method to separate speech into four components: content, pitch, rhythm, and timbre. This method enables recent TTS models to generate speech with various audio styles [25].

Moreover, style transfer can be utilized for adversarial example generation. For example, StyleFool [12] presents an attack based on style transfer to fool video classification models. Jin et al. [28] utilize style-transferred speech audios to attack speech recognition models. SMACK [51] successfully compromise speech recognition models by modifying the prosodies of benign speech samples. In this work, we apply audio style transfer to protect human speech against voice deepfake attacks without introducing additional noise.

7.3 Reverberation Applications

Reverberation is caused by sound waves reflecting off surfaces in an environment. Reverberation effects are particularly noticeable in large indoor environments, where the sound persists after the original sound stops. Haas [23] conducts a comprehensive study on the impact of reverberation on human hearing, noting that humans cannot distinguish reverberations that diminish quickly. Although reverberation is usually very weak, it is not negligible in audio signal processing. A data augmentation approach based on reverberation simulation is used for speech [29] or speaker recognition models [31], improving recognition accuracy, especially in far-field scenarios. Meanwhile, reverberation plays a crucial role in over-the-air adversarial audio attacks. Since adversarial speech samples are created by adding subtle perturbations to the original audio, which are weak and easily affected by reverberation, these attacks often fail to compromise speech recognition models in the physical world. To overcome this weakness, Imperio [41] designs a solution to estimate the RIR in the room and adjust the initial adversarial audio signal, while Chen et al. [16] further improve the effective attack range through channel state estimation. Recent work points out that reverberation can also be used for adversarial audio signal generation. For example, AdvReverb [14] designs an adversarial audio attack by adding reverberation, and TrojanRoom [15] proposes a new attack using RIR to trigger a backdoor in speech recognition without injecting extra noise. In ClearMask, we employ sophisticated manipulation of natural RIR to generate reverberation in speech, providing protection against voice deepfake attacks while preserving speech naturalness and quality.

8 Conclusion

The misuse of voice synthesis technology presents a significant threat to voice data security and privacy. Although numerous defense mechanisms have been proposed in previous studies, they often demonstrate limited effectiveness in real-world applications. In this paper, we introduce ClearMask, a noise-free defense method designed to mitigate voice deepfake attacks by unknown adversaries. Additionally, we propose LiveMask, a real-time protection mode of ClearMask, which is intended for instant communication applications. We evaluate ClearMask and LiveMask against various unseen open-source and commercial voice cloning models. The experimental results demonstrate that ClearMask and LiveMask can effectively prevent synthesized voice from deceiving ASV models or human ears while preserving speech clarity and naturalness. In summary, ClearMask is the first defense mechanism to successfully integrate effectiveness, transferability, naturalness, and real-time capability in combating deepfake voice generation.

Acknowledgments

We would like to extend our appreciation to the shepherd and anonymous reviewers for their invaluable input on our study. This work was supported in part by the U.S. National Science Foundation grant CNS-2310207.

References

[1] 2010. Apple FaceTime. https://apps.apple.com/us/app/facetime/.

[2] 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.

[3] 2023. ElevenLabs. https://www.elevenlabs.io/.

[4] 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.

[5] 2023. plat.ht. https://play.ht/.

[6] 2023. 'Mom, these bad men have me': She believes scammers cloned her daughter's voice in a fake kidnapping. https://www.cnn.com/2023/04/29/us/ai-scam-calls-kidnapping-cec/index.html.

[7] 2024. ChatGPT 4.0. https://chatgpt.com/.

[8] 2024. Google Text-to-Speech. https://cloud.google.com/text-to-speech.

[9] 2024. Soniox: Introducing AudioMind. https://soniox.com/.

[10] Muhammad Ejaz Ahmed, Il-Youp Kwak, Jun Ho Huh, Iljoo Kim, Taekkyung Oh, and Hyoungshick Kim. 2020. Void: A fast and light voice liveness detection system. In *29th USENIX Security Symposium (USENIX Security 20)*. 2685–2702.

[11] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670* (2019).

[12] Yuxin Cao, Xi Xiao, Ruoxi Sun, Derui Wang, Minhui Xue, and Sheng Wen. 2023. Stylefool: Fooling video classification systems via style transfer. In *2023 IEEE symposium on security and privacy (SP)*. IEEE, 1631–1648.

[13] Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*. PMLR, 2709–2720.

[14] Meng Chen, Li Lu, Jiadi Yu, Zhongjie Ba, Feng Lin, and Kui Ren. 2023. AdvReverb: Rethinking the Stealthiness of Audio Adversarial Examples to Human Perception. *IEEE Transactions on Information Forensics and Security* (2023).

[15] Meng Chen, Xiangyu Xu, Li Lu, Zhongjie Ba, Feng Lin, and Kui Ren. 2024. Devil in the Room: Triggering Audio Backdoors in the Physical World. In *33th USENIX security symposium (USENIX Security 24)*.

[16] Tao Chen, Longfei Shangguan, Zhenjiang Li, and Kyle Jamieson. 2020. Metamorph: Injecting inaudible commands into over-the-air voice controlled systems. In *Network and Distributed Systems Security (NDSS) Symposium*.

[17] Yining Chen, Min Chu, Eric Chang, Jia Liu, and Runsheng Liu. 2003. Voice conversion with smoothed GMM and MAP adaptation. In *Eighth European Conference on Speech Communication and Technology*.

[18] Yen-Hao Chen, Da-Yi Wu, Tsung-Han Wu, and Hung-yi Lee. 2021. Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5954–5958.

[19] Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee. 2019. One-shot voice conversion by separating speaker and content representations with instance normalization. *arXiv preprint arXiv:1904.05742* (2019).

[20] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143* (2020).

[21] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2414–2423.

[22] Eric Grinstein, Ngoc QK Duong, Alexey Ozerov, and Patrick Pérez. 2018. Audio style transfer. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 586–590.

[23] Helmut Haas. 1972. The influence of a single echo on the audibility of speech. *Journal of the audio engineering society* 20, 2 (1972), 146–159.

[24] Chien-yu Huang, Yist Y Lin, Hung-yi Lee, and Lin-shan Lee. 2021. Defending your voice: Adversarial attack on voice conversion. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 552–559.

[25] Rongjie Huang, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. 2022. Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech. *Advances in Neural Information Processing Systems* 35 (2022), 10970–10983.

[26] Shehzeen Hussain, Paarth Neekhara, Shlomo Dubnov, Julian McAuley, and Farinaz Koushanfar. 2021. {WaveGuard}: Understanding and mitigating audio adversarial examples. In *30th USENIX security symposium (USENIX Security 21)*. 2273–2290.

[27] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems* 31 (2018).

[28] Weifei Jin, Yuxin Cao, Junjie Su, Qi Shen, Kai Ye, Derui Wang, Jie Hao, and Ziyao Liu. 2024. Towards Evaluating the Robustness of Automatic Speech Recognition Systems via Audio Style Transfer. In *Proceedings of the 2nd ACM Workshop on Secure and Trustworthy Deep Learning Systems*. 47–55.

[29] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. 2017. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5220–5224.

[30] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems* 33 (2020), 17022–17033.

[31] Weiwei Lin and Man-Wai Mak. 2022. Robust speaker verification using population-based data augmentation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7642–7646.

[32] Zihao Liu, Yan Zhang, and Chenglin Miao. 2023. Protecting Your Voice from Speech Synthesis Attacks. In *Proceedings of the 39th Annual Computer Security Applications Conference*. 394–408.

[33] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

[34] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[36] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Kudinov, and Jiansheng Wei. 2021. Diffusion-based voice conversion with fast maximum likelihood sampling scheme. *arXiv preprint arXiv:2109.13821* (2021).

[37] Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox. 2020. Unsupervised speech decomposition via triple information bottleneck. In *International Conference on Machine Learning*. PMLR, 7836–7846.

[38] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*. PMLR, 5210–5219.

[39] Marco A Martínez Ramírez, Oliver Wang, Paris Smaragdis, and Nicholas J Bryan. 2021. Differentiable signal processing with black-box audio effects. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 66–70.

[40] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A General-Purpose Speech Toolkit. arXiv:2106.04624 [eess.AS] arXiv:2106.04624.

[41] Lea Schönherr, Thorsten Eisenhofer, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. 2020. Imperio: Robust over-the-air adversarial examples for automatic speech recognition systems. In *Proceedings of the 36th Annual Computer Security Applications Conference*. 843–855.

[42] Christian J Steinmetz, Nicholas J Bryan, and Joshua D Reiss. 2022. Style transfer of audio effects with differentiable signal processing. *arXiv preprint arXiv:2207.08759* (2022).

[43] James Traer and Josh H McDermott. 2016. Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences* 113, 48 (2016), E7856–E7865.

[44] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. 2016. Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. (2016).

[45] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4879–4883.

[46] Yuanda Wang, Hanqing Guo, Guangjing Wang, Bocheng Chen, and Qiben Yan. 2023. Vsmask: Defending against voice synthesis attack via real-time predictive perturbation. In *Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks*. 239–250.

[47] Yuanda Wang, Hanqing Guo, and Qiben Yan. 2022. GhostTalk: Interactive Attack on Smartphone Voice System Through Power Line. In *Network and Distributed Systems Security (NDSS) Symposium*.

[48] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International conference on machine learning*. PMLR, 5180–5189.

[49] Yuanda Wang, Qiben Yan, Nikolay Ivanov, and Xun Chen. 2023. A Practical Survey on Emerging Threats from AI-driven Voice Attacks: How Vulnerable are Commercial Voice Control Systems? *arXiv preprint arXiv:2312.06010* (2023).

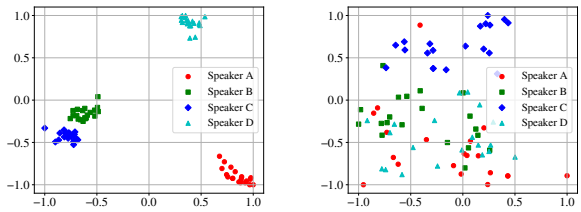
[50] Emily Wenger, Max Bronckers, Christian Cianfarani, Jenna Cryan, Angela Sha, Haitao Zheng, and Ben Y Zhao. 2021. " Hello, It's Me": Deep Learning-based Speech Synthesis Attacks in the Real World. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 235–251.

[51] Zhiyuan Yu, Yuanhaur Chang, Ning Zhang, and Chaowei Xiao. 2023. {SMACK}: Semantically Meaningful Adversarial Audio Attack. In *32nd USENIX Security Symposium (USENIX Security 23)*. 3799–3816.
[52] 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*. 460–474.
[53] 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/.

Appendix

A Voice Feature and Audio Style

When using audio style transfer to process speech audio, the most critical challenge is: *how to select the target audio style for protection?* A straightforward solution is to select an audio style from another speaker whose voice is different from the protected speaker, and apply it to the original speech. However, this is impractical. Fig. 9 shows the distributions of voice embedding vectors and audio style embedding vectors from the same speech samples but different speakers, with all vectors are reduced in dimensionality using principal component analysis (PCA). The voice embedding vectors exhibit clear boundaries between different speakers, while the audio style embeddings from different speakers are randomly mixed. This is because voice encoders are only trained on human speech, whereas audio style features are derived from a wide range of sounds beyond human language, covering a broader spectrum of auditory characteristics. As a result, audio style extraction cannot precisely distinguish between voice characteristics, suggesting that different voices may share similar audio styles. This insight motivates us to explore new audio styles to generate effective adversarial examples.



(a) Voice Embedding Distribution (b) Style Embedding Distribution

Figure 9: Voice and audio style embedding vectors from the same speakers.

B Textual Input Samples

Table 8 shows a few textual input samples for deepfake voice synthesis, which are generated via GPT-4.

Table 8: Textual input samples for deepfake voice synthesis.

Sentences
Hi, I'm calling from HR. We're updating our records and need your Social Security number.
This is the medical office. We require the patient's file to be sent to a new email immediately.
Finance department here. We need the wire transfer code for the recent transaction to finalize it.
This is tech support. We've detected unusual activity on your account. Please provide your password for verification.
I'm outside the building without my ID. Can you buzz me in or give the access code?
This is from IT support. We've noticed a security breach. Please provide your username and temporary password immediately.
Hi, we're updating the security protocol. Can you confirm your employee ID and access badge number for verification?
Hello, I'm handling a critical project update. Can you email me the latest financial forecast document right now?
This is the service desk. To restore your account access, we need you to confirm your mother's maiden name and birth date.
Calling from customer service. To prevent your account from being locked, please provide the recent one-time password sent to you.