VSMask: Defending Against Voice Synthesis Attack via Real-Time Predictive Perturbation

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

Hanqing Guo
Michigan State University
East Lansing, Michigan, USA
guohanqi@msu.edu

Guangjing Wang
Michigan State University
East Lansing, Michigan, USA
wanggu22@msu.edu

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

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

ABSTRACT

Deep learning based voice synthesis technology generates artificial human-like speeches, which has been used in deepfakes or identity theft attacks. Existing defense mechanisms inject subtle adversarial perturbations into the raw speech audios to mislead the voice synthesis models. However, optimizing the adversarial perturbation not only consumes substantial computation time, but it also requires the availability of entire speech. Therefore, they are not suitable for protecting live speech streams, such as voice messages or online meetings. In this paper, we propose VSMask, a real-time protection mechanism against voice synthesis attacks. Different from offline protection schemes, VSMask leverages a predictive neural network to forecast the most effective perturbation for the upcoming streaming speech. VSMask introduces a universal perturbation tailored for arbitrary speech input to shield a real-time speech in its entirety. To minimize the audio distortion within the protected speech, we implement a weight-based perturbation constraint to reduce the perceptibility of the added perturbation. We comprehensively evaluate VSMask protection performance under different scenarios. The experimental results indicate that VSMask can effectively defend against 3 popular voice synthesis models. None of the synthetic voice could deceive the speaker verification models or human ears with VSMask protection. In a physical world experiment, we demonstrate that VSMask successfully safeguards the real-time speech by injecting the perturbation over the air.

CCS CONCEPTS

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

KEYWORDS

Voice Synthesis, Real-time Adversarial Machine Learning, Privacy Enhancement

1 INTRODUCTION

The recent development in Artificial Intelligence (AI) has spurred the research on intelligent voice control. It is now possible to impersonate someone's voice almost impeccably, as shown in various "deepfake" campaigns for misinformation and fraud [1]. Deep learning based voice synthesis is the core technology used in deepfakes to generate speeches of arbitrary target speaker's voice without specific speech samples. There are two types of speech synthesis approaches: text-to-speech (TTS) synthesis and voice conversion (VC). TTS generates synthetic speeches using the script texts and a single speech sample from the speaker. However, it usually outputs unnatural and mechanical human speeches. VC, on the other hand, directly transforms the voice of the original speaker to that of another speaker without altering the linguistic information within the speech. VC is capable of generating vivid speech with real human speech characteristics, such as pauses, moods, and even accents.

Advanced voice synthesis has been used in benign applications, such as reading stories to kids with their deceased grandma's voice [2], or letting your favorite movie star to guide your routes [3]. However, as shown in Fig. 1, human ears or automatic speaker verification (ASV) systems verify speaker identity through voice patterns, voice synthesis technology causes potential security threats. First, voice assistants often require the authenticated voice to activate Siri or Google Assistant before users operate the locked smartphone. Attackers could easily take control of the smartphone and send malicious commands with synthesized voice. Second, ASV systems have been integrated in many popular Apps for user authentication. For example, WeChat requires the user to speak a 6-digit number before the first-time login on an unseen device [4]. Such user verification is vulnerable to voice synthesis attacks. Third, attackers can leverage synthetic voice to impersonate the victim in a fraud phone call [5]. It is difficult for human ears to differentiate high-quality synthetic speech from natural human voice [6].

In response to the threat of voice synthesis attacks, researchers have developed synthetic human voice detection models to identify
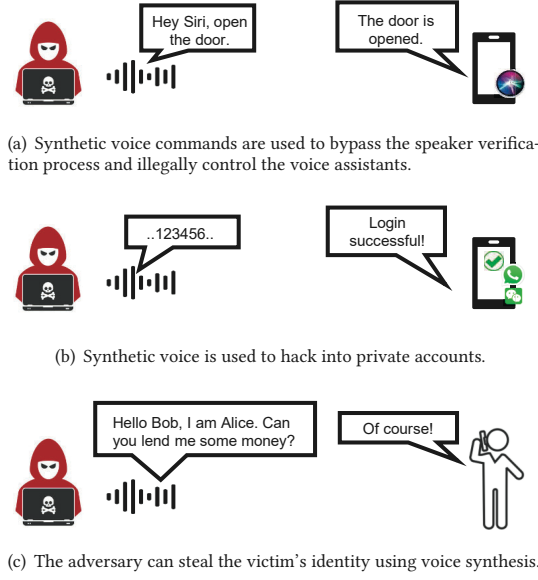
Yuanda Wang, Hanqing Guo, Guangjing Wang, Bocheng Chen, & Qiben Yan

(a) Synthetic voice commands are used to bypass the speaker verification process and illegally control the voice assistants.

(b) Synthetic voice is used to hack into private accounts.

(c) The adversary can steal the victim's identity using voice synthesis.

Figure 1: Voice synthesis attack deceives both ASV system and human ears in real-world scenarios.

fake human speech [7–9]. However, implementing voice synthesis detection on all mobile devices can be costly, and these detection methods may not achieve the claimed high accuracy in real-world scenarios [6]. To defend against unauthorized voice synthesis attacks, Attack-VC [10] adds adversarial perturbations to victim's speech samples to disallow adversaries from generating synthetic voices from the victim. The synthesized speech derived from the protected speech will sound like a distinct voice from another pre-determined target speaker rather than the victim speaker[1]. Nevertheless, the application of Attack-VC is somewhat restricted. The PGD method used to optimize the adversarial perturbation requires the availability of the entire speech and a significant number of iterations to generate the adversarial speech samples. Therefore, the optimization process will not be able to keep up with the real-time speech. As a result, live human speech and voice signals in real-time applications, such as voice messages, online meetings, are still vulnerable to voice synthesis threats.

In this work, we propose the *first* real-time voice protection method, VSMask, to protect the real-time speech in both digital and physical world. Instead of optimizing the perturbation itself, we train a neural network to anticipate the most efficient perturbations in the future. As a result, the protective perturbation can be instantly incorporated into the real-time audio stream without causing extra time delay. In addition, to protect the entire speech, we introduce universal perturbation headers to mask the beginning of the speech regardless of the speech content. Based on human auditory perception characteristics, we introduce a supplementary weight-based amplitude constrained method to further alleviate the audio distortion of the protected speech. In summary, the paper makes following contributions:

- We propose VSMask, a real-time protection system against voice synthesis attacks. By predicting the perturbation for

---

[1]"victim speaker" refers to the target of voice synthesis attack. "target speaker" refers to a different speaker pre-determined by VSMask to interfere with the voice synthesis.
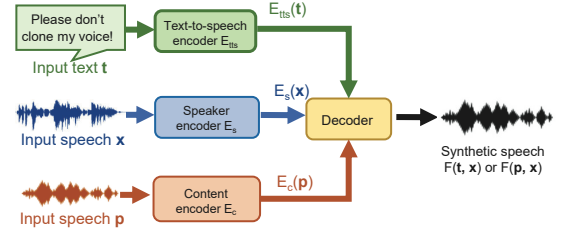

Figure 2: The speaker encoder in voice synthesis models is susceptible to adversarial attacks.

upcoming streaming speech, VSMask can shield the user's voice with negligible delays.
- We propose an optimization scheme to train universal perturbation to protect the exposed audio at the beginning of speech. This allows VSMask to indiscriminately protect speech data with different lengths and sizes. We further introduce a weight-based amplitude constrained method to reduce human perceptibility for adversarial perturbations.
- We evaluate VSMask performance on three different voice synthesis models. The experimental results show that VS-Mask is capable of providing real-time defense against voice synthesis attacks on both digital and physical spaces, effectively shielding ASV systems and the human auditory system from malicious voice synthesis.

2 BACKGROUND
2.1 Voice Synthesis
The purpose of voice synthesis attacks is to generate artificial speech resembling real human voices. Fig. 2 shows a general voice synthesis framework including both TTS and VC. For TTS and VC models, the speech content embeddings are respectively extracted by the content encoders E_{tts} or E_c from text $\mathbf{t}$ or speech $\mathbf{p}$. The speaker encoder E_s encodes the victim speaker characteristics $E_s(\mathbf{x})$, where $\mathbf{x}$ is a speech sample from the victim speaker x. Finally, the decoder will generate a synthetic speech $F(\mathbf{x}, \mathbf{p})$ or $F(\mathbf{x}, \mathbf{t})$ with voice features from an arbitrary victim speaker x.

2.2 Defense Against Voice Synthesis
Attack-VC [10] introduced *embedding attack* that aims at the speaker encoder model E_s. If we input an adversarial speech sample $\mathbf{x} + \delta$ with little difference from $\mathbf{x}$, the encoded vector $E_s(\mathbf{x} + \delta)$ is close to another target speaker y rather than the real victim speaker x. The adversarial attack for voice protection can be formulated in Problem (1) as follows:

$$\begin{aligned} \underset{\delta}{\text{minimize}} \quad & \mathcal{L}(E_s(\mathbf{x} + \delta), E_s(\mathbf{y})) - \lambda \mathcal{L}(E_s(\mathbf{x} + \delta), E_s(\mathbf{x})) \\ \text{subject to} \quad & \|\delta\|_\infty < \epsilon, \end{aligned} \tag{1}$$

where $\mathcal{L}$ is mean squared error (MSE) loss function, and $\mathbf{y}$ is a speech sample from target speaker y. The first term in Problem (1) minimizes the loss between the speaker embedding vector of adversarial input $E_s(\mathbf{x} + \delta)$ and the target speaker $E_s(\mathbf{y})$ to force the model to generate speech voice similar to the target speaker y, while the second term eliminates the victim speaker voice $E_s(\mathbf{x})$. Eventually, with protected speech input $x + \delta$, the synthetic voice cannot spoof the speaker verification or human ears anymore. In Problem (1), it is necessary to set a predetermined target speaker y. We run an additional experiment without setting a target speaker,
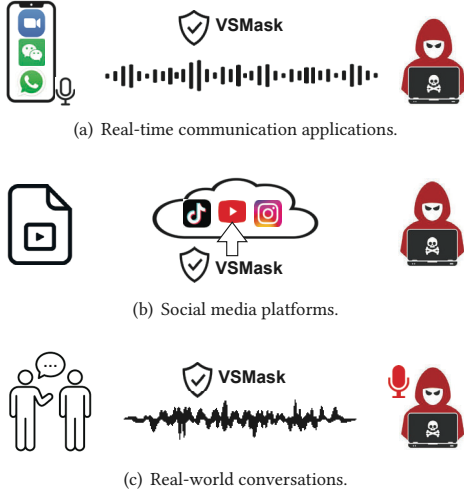
(a) Real-time communication applications.

(b) Social media platforms.

(c) Real-world conversations.

Figure 3: VSMask provides protection in both online and offline scenarios.

(i.e., by setting $\lambda \to \infty$). The result shows: despite that the synthesized speech becomes noisy, it still sounds like the voice from the victim speaker.

2.3 Real-time Adversarial Attack

Since human speech is a streaming audio signal, it is impractical to generate real-time adversarial perturbation using the PGD method. VoiceCamo [11] presents a real-time adversarial attack towards speech recognition (SR) models, which could forecast the upcoming adversarial perturbation by optimizing the model g_θ. The predictive model optimization in VoiceCamo is illustrated as follows:

$$\underset{\theta}{\text{maximize}} \ \mathbb{E}_{(x_t, y_t)}[\mathcal{L}_{CTC}(\bar{y}_t, y_t)]$$
$$\textbf{s.t.} \quad \bar{y}_t = f_\psi(x_t + g_\theta(x_{t-r-\delta})) \quad \text{and} \quad \|g_\theta(x_t)\|_\infty < \epsilon, \tag{2}$$

where $\mathcal{L}_{CTC}$ is connectionist temporal classification (CTC) loss function, x_t is a streaming speech audio, and y_t is the true label of the speech transcript. f_ψ is the speech-to-text model with *fixed* parameter ψ, g_θ is the predictive model to forecast the perturbation, and $x_{t-r-\delta}$ is the speech signal captured at time $t - r - \delta$. After predicting the first perturbation, the input window slides to the next time slot to predict the second perturbation, and so on so forth. Finally, the speech recognition accuracy will be significantly degraded when the real-time attack activates.

3 THREAT MODEL

The state-of-the-art voice synthesis models allow the attackers to generate voice with a single speech sample from the victim speaker. In this section, we define the threat model of voice synthesis attacks according to the adversary's knowledge about the victim and potential defenses.

3.1 Adversary Knowledge

The attacker can leverage voice synthesis models to generate artificial speech with the victim speaker's voice, in order to compromise the speaker verification systems or human ears. The model could be well-trained or locally trained by the adversaries. They can also test the synthesized speech on speaker recognition models to evaluate the voice similarity. If they realize the samples are protected, they

can apply denoising or other signal processing methods to recover the clear audio.

3.2 Adversary Capability

The adversary is able to save audio signals played in mobile devices, for example, smartphones or laptops, or record speech in physical world with a microphone. The victim speaker's speech samples are available from multiple sources.

Real-time Communication Applications. As shown in Fig. 3(a), it is common for people to send voice memos or make voice calls using real-time communication apps. The attackers may hide in chat groups and record anyone's voice for voice synthesis. Meanwhile, many people meet online since the COVID-19 pandemic. The attackers can join public online meetings and record the victim's voice without seeking permission. But existing defense methods fail to provide protection for the real-time speech.

Social Network Platforms. People are willing to share their daily life on social media. There are billions of active users uploading their self-made videos to TikTok, YouTube and Instagram [12]. Usually, all visitors have the access to download and reuse them. Even with a short speech, malicious attackers could clone your voice and impersonate your identity. Although existing defense methods are able to protect your voice on your social media, it requires long processing time before uploading [10].

Real-world Conversation. In addition, real-world human speech is threatened by voice synthesis attacks. While people are speaking in a presentation or conversation, an adversary can stealthily record the speech and launch voice synthesis attacks. But existing defenses fail to mask our voice efficiently to allow the protection against malicious voice synthesis. It is also challenging to physically inject perturbation into real speech.

3.3 Defender Knowledge

Meanwhile, we consider a white-box scenario where the defenders have complete knowledge of the adversarial voice synthesis model, including model structure and model parameters. To optimize VSMask, defenders also need to collect a few clear speech samples from the protected victim speaker x. In addition, defenders have multiple target speaker candidates y_i speech samples from public datasets, and they can select the best target y for the victim speaker x to maximize the voice disparity. Raw speech audios will be locally processed by VSMask and uploaded to real-time applications or social media platforms without extra time delay.

4 VSMASK SYSTEM DESIGN

4.1 Design Consideration

One typical attack approach towards ASR system is to hide the adversarial perturbations in the psycho-acoustic features of human speech [13], which could be a potential solution for voice synthesis defense. However, these perturbations require the availability of the entire speech to mask them. As a result, it cannot be applied in real-time applications.

We also notice that universal perturbations can fool the ASR system regardless of the original speech contents [14]. But such short perturbations are too short to protect the long speech from voice synthesis attacks. To extend the protection scope, we attempt to leverage periodical universal perturbation to shield human speech. Unfortunately, the performance of periodical perturbation method
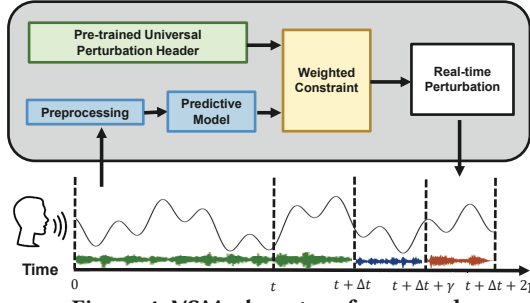
Yuanda Wang, Hanqing Guo, Guangjing Wang, Bocheng Chen, & Qiben Yan

Figure 4: VSMask system framework.

is unsatisfying. The defense is effective only when the perturbation signal δ is strong. Generally, short universal perturbations have limited degrees of freedom, which limits the protection performance.z In contrast, longer perturbation period usually causes over-fitting; besides, periodical perturbation can be denoised by attackers once they capture the perturbation at the interval of the speech. Thus, it is crucial to develop a robust and real-time protection mechanism to safeguard against voice synthesis attacks.

4.2 Real-time Predictive Model

Preprocessing. Fig. 4 shows the system framework of VSMask. The input of VSMask is a streaming human speech signal, which is vulnerable to voice synthesis attacks. Once VSMask detects human speech starts, it will capture the upcoming audio signal. During preprocessing, the audio signal is transformed into a mel-spectrogram, with the frequencies converted to the non-linear mel scale to match human hearing. The resulting mel-spectrogram has the same dimensions as the input of the speaker encoder E_s in the targeted voice synthesis model.

Model Architecture. Next, the converted mel-spectrogram works as the input of VSMask predictive model f_θ. The output of the predictive model is also a mel-spectrogram with shorter length. Learned from VoiceCamo [11], VSMask selects a real-time predictive model composed of 7 down-sampling blocks and 5 up-sampling blocks, which outperforms models with less layers and maintains similar performance as models with more layers. Each down-sampling block contains a reflection padding followed by a 2-D convolution layer, a 2-D batch form, and a PReLU activation function; each up-sampling block consists of a 2-D ConvTranspose layer and a leaky ReLU activation function. The final up-sampling block is followed by a tanh function to constrain $\|\delta\|_\infty$. The output mel-spectrogram will be converted to time-domain audio signal and injected into the raw audio. We illustrate the fine-tuning process and model architecture details in Appendix A.

Optimization and Workflow. Suppose the speaker encoder function in the target voice synthesis model is E_s, we could formulate predictive model training as the following optimization problem:

$$\begin{aligned}
\underset{\theta}{\text{minimize}} \quad & \mathcal{L}(E_s(\mathbf{x}+\delta), E_s(\mathbf{y})) - \lambda\mathcal{L}(E_s(\mathbf{x}+\delta), E_s(\mathbf{x})), \\
\text{subject to} \quad & \delta^\gamma_{t+\Delta t} = f_\theta(\mathbf{x}_t) \quad \text{and} \quad \|\delta\|_\infty < \epsilon,
\end{aligned} \tag{3}$$

where $\mathcal{L}$ is MSE loss function, $\mathbf{x}$ is the input streaming speech from the victim speaker x, $\mathbf{y}$ is a speech sample from the selected target speaker y to mislead E_s, and δ is the perturbation predicted by model f_θ. $\delta^\gamma_{t+\Delta t}$ is the perturbation added at $t + \Delta t$ with length γ. Since the predictive model f requires an input signal $\mathbf{x}_t$ with a short

length t, VSMask does not require the entire speech in advance, facilitating real-time protection.

As illustrated in Figure 4, VSMask initially generates a perturbation of length γ (represented by the blue waveform) after a short time delay of Δt. Subsequently, the input window is slid to the next time slot from γ to $t + \gamma$ to generate the second perturbation (represented by the red waveform), and this process repeats. In contrast to Attack-VC, VSMask fine-tunes the predictive model parameters θ rather than manipulating the perturbation δ. As a result, once the predictive model f_θ is established, VSMask can identify the best perturbation δ in a single computation step, rather than undergoing a large number of iterations.

4.3 Universal Perturbation Header

However, VSMask predictive model has some deficiencies, since the voice synthesis models can generate synthetic speech via very short speech samples from the victim speaker. For example, the victim speaker gives a speech example – *"please protect my speech from voice synthesis"*. The predictive model could well protect the trailing phrases *"my speech from voice synthesis"*. Yet, the leading words of the speech *"please protect"* are still exposed to the attackers, as VSMask prediction model cannot generate perturbation for the beginning of the speech. Therefore, the attacker may be able to clip the speech audio and synthesize speech with an audio segment.

To address this challenge, we resort to a universal perturbation added at the beginning of the streaming speech signal to protect the entire speech. Recent work [14] shows that regardless of the utterance, there exists universal perturbations that can force the ASR system to output specific transcripts, e.g., "yes" or "no". This universal perturbation maintains high success rate especially when the input speech is short. Fortunately, the "blind spot" of the predictive model is also small. Therefore, it is feasible to train a universal perturbation forcing the speaker encoder to output the selected target speaker y embedded vector $E_s(\mathbf{y})$ rather than the real vector $E_s(\mathbf{x})$, regardless of the speech contents.

In order to train such a universal perturbation header δ_h, we clip the training speech samples from the victim speaker x to audio clips $\mathbf{x}_i$ with the same length $t + \Delta t$. Then, we use PGD method to optimize the universal perturbation header δ_h as follows:

$$\begin{aligned}
\underset{\delta_h}{\text{minimize}} \quad & \sum_{\mathbf{x}_i \in \mathcal{D}} \mathcal{L}_s(\delta_h, \mathbf{x}_i, \mathbf{y}), \\
\text{subject to} \quad & \mathcal{L}_s(\delta_h, \mathbf{x}_i, \mathbf{y}) = \mathcal{L}(E_s(\mathbf{x}_i + \delta_h), E_s(\mathbf{y})) \\
& - \lambda\mathcal{L}(E_s(\mathbf{x}_i + \delta_h), E_s(\mathbf{x}_i)) \quad \text{and} \quad \|\delta_h\|_\infty < \epsilon.
\end{aligned} \tag{4}$$

where $\mathcal{D}$ is the collection of clipped speech $\mathbf{x}_i$.

Before the optimization process in Section 4.2, we add the universal header into all training samples to enhance the protection performance. After combining the predictive model and universal header, VSMask successfully protects the streaming speech in real-time scenarios. Fig. 5(a) shows a natural speech sample $\mathbf{x}$ exposed to the attackers. Fig. 5(b) is the audio spectrogram of protected speech $\mathbf{x} + \delta$. The perturbation causes little difference between these two speech samples. Fig. 5(c) and Fig. 5(d) are respectively the synthetic speech spectrograms from $\mathbf{x}$ and $\mathbf{x}+\delta$ with the same speech content. Even though $\mathbf{x}$ and $\mathbf{x} + \delta$ are very similar, the synthetic speech $F(\mathbf{x})$ shows male voice patterns whereas $F(\mathbf{x} + \delta)$ presents female voice with higher frequencies.
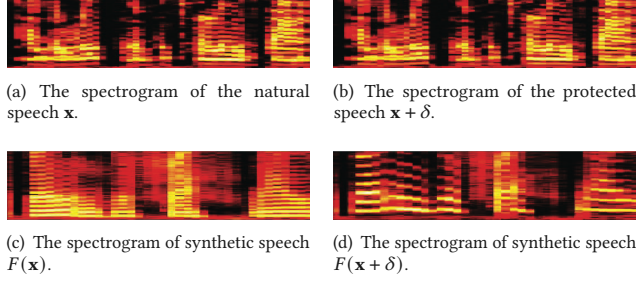
(a) The spectrogram of the natural speech $\mathbf{x}$.

(b) The spectrogram of the protected speech $\mathbf{x} + \delta$.

(c) The spectrogram of synthetic speech $F(\mathbf{x})$.

(d) The spectrogram of synthetic speech $F(\mathbf{x} + \delta)$.

Figure 5: Speech protected by VSMask effectively degrades the voice synthesis model performance.

Figure 6: Equal-loudness contour.

4.4 Weight-based Noise Mitigation

Inevitably, the injected protective perturbation will cause audio distortion. Prior work usually constrains the perturbation amplitude to minimize the human perception, or considers the amplitude in the optimization to achieve the best trade off between noise level and performance. However, even as the perturbation is well-optimized, there is still significant background noise in the speech signal.

To degrade the perturbation perceptibility of VSMask, we look into human hearing sensitivity. Fig. 6 shows equal-loudness contour [15], a measurement of human hearing perception under different frequency tones. For low frequency (< 1.6 kHz) or high frequency (> 4 kHz), higher sound pressure is required to cause the same loudness level in human ears. For medium frequency band between 1.6 kHz and 4 kHz, human ears have the highest sensitivity. Based on this phenomenon, one potential approach is to hide all perturbations in the low frequency band. However, our experiment shows that it fails to provide sufficient protection for our speeches.

Therefore, we need to inject a global perturbation into the entire spectrogram. To mitigate human hearing perception, we are supposed to reduce the perturbation amplitude between 1.6 kHz and 4 kHz. However, it is expected that reducing the perturbation amplitude also degrades the protection performance. To compensate the scarcity in medium frequency, we increase the amplitudes in other frequency bands which are hard to sense for human ears. Therefore, we rewrite the amplitude constraint as follows:

$$\delta = \begin{bmatrix} \delta_l & \delta_m & \delta_h \end{bmatrix}^T,$$
$$\mathbf{s.t.} \quad \|\delta_l\|_\infty < \epsilon_1, \ \|\delta_m\|_\infty < \epsilon_2, \ \|\delta_h\|_\infty < \epsilon_3, \tag{5}$$

where $\epsilon_2 < \epsilon_3 < \epsilon_1$. By adding such weight-based constraint on the perturbation amplitude, we successfully improve the protected audio quality by limiting the human perception of perturbation without sacrificing the protection performance.

4.5 VSMask Application Scenarios

As we mentioned in Section 3, the attackers have many different approaches to obtain the speech audio from the victim speaker. Here, we describe the versatile application scenarios of VSMask.

Online Protection. The first application is to protect human voice in real-time communication and social media Apps. For these Apps, VSMask can work as a plug-in component to protect the user's voice. Once VSMask detects the start of the speech, it will automatically add the universal perturbation header, and concurrently capture the speech audio. After injecting the perturbation header, VSMask has predicted the following perturbation for the next time slot, and then it iteratively protects the upcoming speech stream. Based on
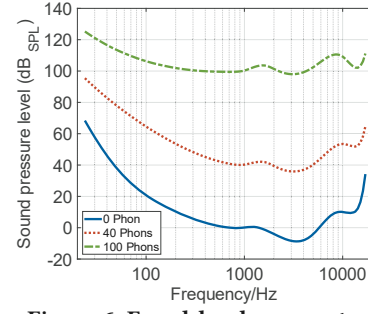
the prediction mechanism, VSMask causes no extra time delay on the real-time communication.

Real-world Physical Protection. In real-world scenarios, the attackers may stealthily record the victim's voice and implement voice synthesis attack. To counter such adversaries, users can protect themselves by installing VSMask on their mobile devices. While talking, VSMask will play perturbation noise signal and effectively inject adversarial speech samples in the eavesdropped audio. In the end, it will protect our real-world speech against voice synthesis attack, as shown in the experiments in Section 6.

5 EVALUATION

5.1 Target Models and Experiment Design

Target Voice Synthesis Model. Moreover, we select the following three representative voice synthesis models to evaluate VSMask performance:

- AdaIN-VC [16]: an unsupervised one-shot voice conversion model based on instance normalization;
- AutoVC [17]: a new style auto encoder with carefully designed bottleneck, which is the first work achieving zero-shot voice conversion with non-parallel speech data;
- SV2TTS [18]: a TTS model that converts text transcripts to human speech with the target speaker voice.

Dataset. For AdaIN-VC and AutoVC models, we use CSTR VCTK Corpus [19] dataset, a widely used speech dataset in the voice conversion studies, to evaluate the performance of VSMask. VCTK includes speech samples from 109 English speakers with different accents. Each speaker reads up to 400 sentences randomly selected from newspapers or elicitation. For SV2TTS model, we select LibriSpeech dataset [20] for evaluation. In particular, as some speech samples are too short to activate VSMask predictive model, we only use speech audios longer than 3 seconds for training and testing.

5.2 VSMask Effectiveness on ASV System

ASV System for Evaluation. To evaluate VSMask effectiveness of reducing the similarity between the synthetic speech voice and victim speaker's voice, we utilize SpeechBrain [21], a state-of-the-art model to calculate the voice similarity of two speech samples. SpeechBrain speaker verification model is composed of emphasized channel attention, propagation and aggregation in time delay neural network (ECAPA-TDNN) to extract speaker embeddings. The similarity scores are calculated by cosine similarity of embedding vectors of two speech samples. The higher the score is, the greater the possibility that two speech samples are from the same speaker.

Yuanda Wang, Hanqing Guo, Guangjing Wang, Bocheng Chen, & Qiben Yan

Table 1: We compare VSMask defense performance with other online or offline baseline defense methods. VSMask outperforms all other real-time protection methods and achieves similar performance as offline PGD method.

Type	Method	Score		Accuracy		STOI	
Non-synthetic	Raw speech	0.780		98.8%		1.000	
	Protected speech	0.429		68.5%		0.839	

Type	Method	M2M		F2F		M2F		F2M	
		Score	ASR	Score	ASR	Score	ASR	Score	ASR
Synthetic	Raw speech	0.595	91.9%	0.612	93.2%	0.561	88.3%	0.546	86.0%
	Random noise	0.516	86.6%	0.538	89.0%	0.505	84.0%	0.473	81.5%
	Periodical perturbation	0.192	11.0%	0.203	12.5%	0.177	9.8%	0.156	8.6%
	Offline PGD	0.064	0.0%	0.085	0.0%	0.049	0.0%	0.055	0.0%
	VSMask	**0.077**	**0.0%**	**0.104**	**0.0%**	**0.056**	**0.0%**	**0.073**	**0.0%**

M2M: Male-to-Male, F2F: Female-to-Female, M2F: Male-to-Female, F2M: Female-to-Male.

If the score is higher than a threshold k, the speech can pass the verification. In our experiment, we set the default threshold $k = 0.25$, which achieves the best accuracy for clear human speech samples. **VSMask and Baseline Methods Setup.** In this section, we introduce VSMask evaluation setup along with 4 other baseline methods for comparison. We evaluate the methods in the digital domain by injecting perturbations into digital audio signals.

VSMask: First, for VSMask method, we set $\lambda = 1$ in Problem (3) and (4) according to the λ value in [6], the input time window length t as 1.25 seconds. To ensure real-time prediction, it is important to avoid setting a very short time-delay. On the other hand, a long time-delay can lead to a degradation in the performance of VSMask. Therefore, we have set both the time-delay (Δt) and output length (γ) to be 0.4 seconds. Therefore, the perturbation header length is $t + \Delta t = 1.65$ seconds. Moreover, we set the perturbation amplitude constraint $\epsilon = 0.10$. The perturbation amplitude after noise mitigation approach is: $\epsilon_3 = \epsilon$, $\epsilon_1 = 1.15 \epsilon$ and $\epsilon_2 = 0.85 \epsilon$.

In addition, we randomly select 30 male and 30 female speakers from the dataset. We use AdaIN-VC model to generate 100 synthesized speech samples for each speaker. For female victim speakers, we select a male speaker as the target speaker, and vice versa. For voice conversion models, the gender of the speaker matters, since male and female have different voice patterns, e.g., female speakers usually have higher frequency utterances. We split the results according to different genders of victim speaker and speech source speaker.

Random Noise: One possible way to degrade the voice synthesis model performance is to add random noise into the speech signal **x**. In the experiment, we set the random noise amplitude $\epsilon_n = 0.15$.

Periodical Perturbation: Since the universal header can protect short speech regardless of the speech content, periodical universal perturbation is a potential solution for real-time defense. In the experiment, we recurrently inject the same universal perturbation header with amplitude $\epsilon_p = 0.12$ into the raw speech audio and evaluate its protection performance.

Offline PGD Method: In addition, we compare our method with offline PGD method in Attack-VC [10]. We apply the embedding attack method as shown in Problem (1), while the iteration number is $1, 500$, $\epsilon_{off} = 0.10$. Even though offline PGD method is not feasible for real-time protection, it can help to find the upper bound of protection performance.

VSMask Evaluation on ASV System. We quantitatively evaluate the ASV system performance with non-synthetic speech samples. For the clear audios without VSMask protection, the ASV system achieves accuracy above 98%. But for speech audios protected by VSMask, similar to the result in Attack-VC, the ASV accuracy degrades since the perturbation distorts the voice patterns in the speech. In this way, VSMask may affect speaker recognition functions in real-world scenario such as voice assistant activation. Fortunately, mobile devices can cancel out the perturbation from loudspeakers using echo cancellation [22]. So VSMask will not interfere with the activation process. In addition, we use Short-Time Objective Intelligibility (STOI) [23] to measure the speech quality of protected audio. The result shows that the speech under VSMask protection is fully intelligible for human ears.

Next, we compare VSMask with different baseline defense methods, where the target voice synthesis model is AdaIN-VC. We use attack success rate (ASR) to evaluate the performance of voice synthesis attacks. For reference, when the attacker clones voice with raw speech audios, more than 85% synthetic audios can pass the ASV system with high similarity score. While the voice conversion is intra-gender (male-to-male or female-to-female), the ASR is even higher. If the victim speaker is female, voice conversion shows slightly higher ASR than male victim speaker. This result confirms the effectiveness of compromising ASV systems with voice synthesis attacks.

Then, we evaluate the baseline real-time defense methods. We first synthesize voice from speech with white noise. When the perturbation amplitude is small ($\epsilon_n = 0.15$), random noise cannot deteriorate the voice synthesis performance. Actually, the embedding speaker vectors $E_s(\mathbf{x})$ and $E_s(\mathbf{x} + \delta)$ show little difference when δ is weak random noise. If stronger noise is implemented, it works because of poor speech quality. When the noise amplitude $\epsilon > 0.4$, no synthetic speech samples can compromise ASV model.

(a) The similarity scores for different voice synthesis models

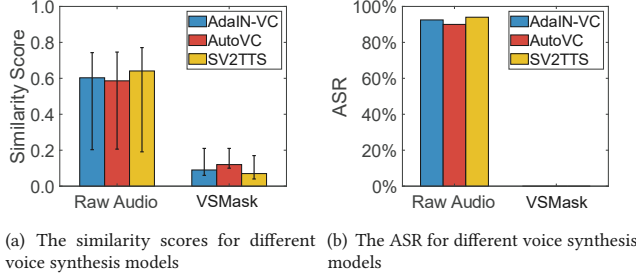(b) The ASR for different voice synthesis models

Figure 7: For all target voice synthesis models, VSMask significantly reduce the voice similarity. With VSMask protection, no synthetic speech can pass the speaker verification system.

But the protected audio $\mathbf{x} + \delta$ is illegible for human ears. Consequently, strong noise is not a feasible protection method in real world.

Second, we use the periodical universal perturbation for defense. As shown in Table 1, periodical universal perturbation significantly reduces the similarity of the synthetic voice and the victim's voice. But there are still more than 10% intra-gender synthetic audios that could pass ASV systems. The reason is that when perturbation amplitude is small, universal perturbation cannot compromise the entire speech, especially for long speech samples. We can increase the amplitude or cycle length to improve the periodical method performance. However, if the perturbation is too strong, it can lead to significant audio distortion, and using a long period for perturbation may not be effective in protecting short speech segments.

Finally, we compare the offline PGD method performance with VSMask. Theoretically, offline PGD method achieves the best performance for defending against voice synthesis attacks. No synthetic audio from offline PGD protected speech can pass ASV systems. Its performance can be further enhanced by increasing the iterations. But for 1,500 iterations, its time consumption is approximately 20× audio length. The computation time could be much longer when the model is deployed on devices without high performance GPU. We mark VSMask performance in **bold** in Table 1. Compared with all other online baseline methods, VSMask has the strongest protective impact. And compared with offline PGD method, VSMask achieves similar offline protection performance without incurring any time delay. For both intra-gender and inter-gender voice conversions, no synthetic speech samples can pass ASV system. Therefore, the results demonstrate that VSMask can successfully protect human speech against unauthorized voice synthesis attacks.

5.3 VSMask Evaluation on Different Voice Synthesis Models

In this section, we evaluate VSMask performance on different voice synthesis models. From Section 5.2, we learn that for voice conversion models, intra-gender voice conversion has better similarity than inter-gender voice conversion. For AutoVC model we generate 50 intra-gender synthesized speech samples for each speaker.

Fig. 7(a) and Fig. 7(b) show the synthetic speech similarity and ASR before VSMask protection. For VC models, after being trained with data in VCTK corpus, both AdaIN-VC and AutoVC synthetic

voice samples can pass ASV with more than 90% success rate. AutoVC shows smaller variance because of its well-designed bottleneck mechanism. But AdaIN-VC has better overall performance on similarity scores. We also test TTS model and compare its performance with VC models. The SV2TTS model is trained with 16 kHz speech samples. To guarantee its best performance, we select 60 victim speakers (30 males and 30 females) from LibriSpeech dataset. Also, we generate 50 synthesized speech samples from each speaker for testing. Then, we test all generated text-to-speech samples in ASV system. Compared with VC models, regardless of the input phrase length, SV2TTS achieves the best similarity score and highest ASR. The potential reason is that SV2TTS generates speech content directly from text input, so that it contains no prior existing voice pattern. But for VC models, even though they are well optimized, there is residual voice pattern from the source speech $\mathbf{p}$. This will deteriorate the voice similarity of the synthetic speech and victim speaker.

In contrast, we also display the similarity score and ASR after VSMask protection. For VC models, VSMask substantially reduces the similarity score between synthetic speech and the raw speech from the victim speaker. None of the synthetic speech samples could compromise ASV system. Since the synthetic voice patterns in SV2TTS are entirely determined by the reference speech $\mathbf{x} + \delta$, this model is more susceptible to VSMask protection compared to VC models. Overall, under VSMask protection, it is unlikely for attackers to spoof speaker verification system by either VC or TTS.

5.4 Human Study

Recent work [6] shows voice synthesis attack not only compromises ASV systems, but also deceives human ears. In this section, we run a human study of VSMask[2]. We select AdaIN-VC and AutoVC for synthetic voice generation because voice conversion models have more natural utterance than TTS.

In the human study evaluation, we recruit 25 anonymous volunteers with normal hearing ability (14 males and 11 females, ages from 20 to 38 years old), and ask them to listen to 12 groups of speech samples through the same loudspeaker with fixed volume (∼60 dBA). 6 groups of synthetic speech samples are from AdaIN-VC model and 6 groups are from AutoVC model. For each model, half of samples are male-to-male and the other half is female-to-female. Each group contains 4 audio samples: ① raw speech audio without protection, ② the same speech audio but with VSMask protection, ③ synthetic voice sample from the raw speech, ④ the same synthetic voice sample from VSMask protected speech. All the speech sample lengths are from 4 seconds to 8 seconds.

First, we ask the volunteers about the noise level in VSMask protected speech. After listening to the raw speech and protected speech, the listeners are suppose to measure the perturbation perceptibility of the protected sample. Options from low to high are: (A) Cannot hear any noise; (B) Can hear noise, but negligible; (C) Can hear noise, but acceptable; (D) Can hear conspicuous noise; (E) Very strong noise, unacceptable.

We collect all answers and present them in Fig. 8(a). The results depict that for both AdaIN-VC and AutoVC, the noise level in the protected voice samples is acceptable for human ears. Around

---

[2]The IRB request has been approved by the university board.

(a) The noise level of protected speech for voice conversion models.

(b) Human evaluation of VSMask protection performance on AdaIN-VC model.

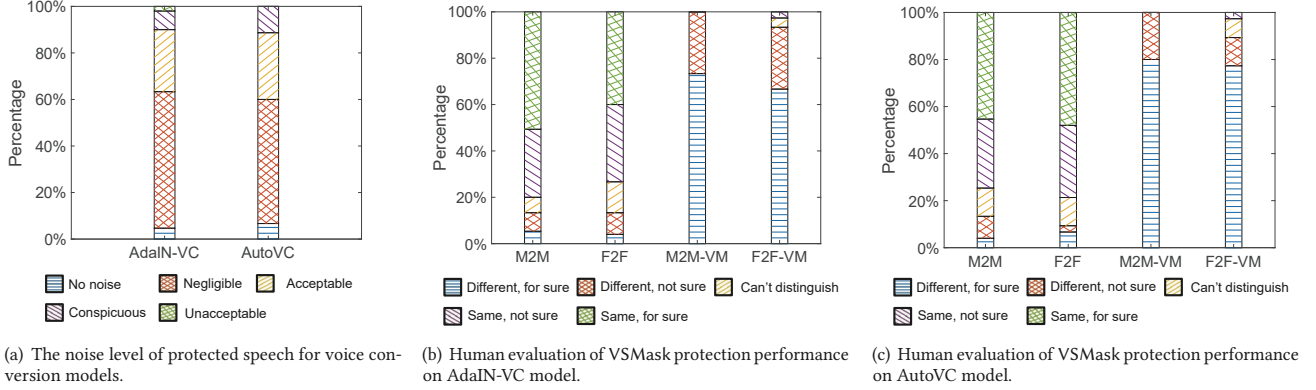(c) Human evaluation of VSMask protection performance on AutoVC model.

Figure 8: Subjective human evaluation results about the protected speech noise level and the protection performance of VSMask.

60% answers are "negligible" or "imperceptible", which means that VSMask protected speech audio has similar clarity as raw audios. More than 90% answers consider the noise in protected audios is acceptable for human ears. Only 2 out of 300 answers point out that the protected speech has "unacceptable" noise.

Next, we ask the volunteers to evaluate the similarities between the raw speech and synthetic speech samples. Similarities options from low to high are: (A) Different voice, for sure; (B) Different voice, but not sure; (C) Cannot distinguish; (D) Same voice, but not sure; (E) Same voice, for sure. Consider the voice conversion performance difference caused by different speaker genders, we separately display results of male-to-male and female-to-female speech samples.

Fig 8(b) and Fig 8(c) show the voice similarity between the ① raw audio & ③ synthetic audio from raw speech, and ① raw audio & ④ synthetic audio from protected speech. For both AdaIN-VC and AutoVC, most synthetic speeches from raw speech can spoof human ears. More than 75% answers indicate that the synthetic speech is definitely or probably from the same speaker as the raw audio. Only 4% answers indicate that the voice is absolutely from another speaker. Therefore, the results show that synthetic voice has sufficient similarity as the victim speaker when launching intra-gender voice conversion. In contrast, for synthetic speech samples from VSMask protected speech (M2M-VM and F2F-VM), they could not fool human ears any more. More than 75% answers believe that the synthetic audios from VSMask protected speeches have definitely different voice from the victim speaker. In total, over 97% answers claim that they will doubt the speaker identity behind the voice. Nobody consider that the synthetic speech matches the victim's voice very well. Furthermore, the offline PGD method has been shown to achieve comparable protection performance in preventing deepfake voices from deceiving human ears [10]. In addition, we present the impact of weight-based noise mitigation in Appendix B.

5.5 Cross-model Evaluation

Attack-VC [10] evaluates defense performance under black-box scenarios by training substitute models. However, sometimes we have no information about which voice synthesis model will be implemented. In this section, we evaluate VSMask performance under cross-model conditions.

Table 2: We generate the adversarial examples from the source model and apply the target model for voice synthesis. The result shows that VSMask can still reduce the ASR under cross-model setup.

Source / Target	AdaIN-VC	AutoVC	SV2TTS
AdaIN-VC	-	15.0%	10.5%
AutoVC	12.8%	-	0.0%
SV2TTS	7.3%	15.2%	-

We apply VSMask to generate adversarial speech samples targeting one voice synthesis model, and apply them on other models. The cross-model ASR is listed in Table 2. Overall, VSMask is still effective for cross-model scenario. But the performance is degraded as the result of different encoder models, preprocessing methods, and training data. For example, AdaIN-VC applies 512-dim mel-spectrogram as speaker encoder input, but 80-dim for AutoVC speaker encoder. Also, SV2TTS is trained with 16 kHz audios, but AdaIN-VC and AutoVC use 48 kHz audios. When we input 16 kHz speech sample in AutoVC, it could not output qualified audios. Moreover, the adversarial samples targeting AdaIN-VC can reduce the similarity of AutoVC synthetic speech as well. When we use adversarial samples from AdaIN-VC or AutoVC on SV2TTS model, a few synthetic samples can still compromise ASV system. We can further apply AdaBelief method [24] to improve the transferability of protected samples by gradually reducing the learning rate, which we leave for future work.

5.6 Adaptive Attack Evaluation

Denoisers are widely deployed in speech enhancement models for audio quality improvement. Adversarial perturbations can also be mitigated by some audio signal transformation methods. To evaluate VSMask robustness against adaptive attackers, we apply denoiser and WaveGuard [25] to remove the perturbation in protected audios and launch voice synthesis by AdaIN-VC model. All synthetic speech samples are from intra-gender speech synthesis.

A simple denoiser fails to remove the well-designed perturbation in protected speech, and the synthetic speech similarity is even further degraded since the denoiser eliminates some low-power signals in the speech. Typically, speech-to-text models can still accurately

Table 3: VSMask performance when adaptive attackers apply denoiser and WaveGuard.

Adaptive methods	None	Denoiser	WaveGuard			
			Down-Up (f=24k)	Quan.-Dequan.	Mel. (Bin=128)	LPC (Ord.=10)
Score	0.096	0.090	0.078	0.082	0.080	0.073

transcribe speech with lower resolution. This allows WaveGuard to compromise speech audio quality while mitigating the adversarial perturbation. We evaluated four WaveGuard signal transformation methods, including Downsampling-Upsampling with different frequencies, Quantization-Dequantization, Mel-spectrogram conversion with different mel bins, and Linear Predictive Coding (LPC) with different orders, on speech samples protected by VSMask. We list the best similarity score in Table 3. In contrast to adversarial examples that target speech-to-text models, WaveGuard methods cannot undermine the protection performance of VSMask since VSMask optimizes the perturbation in mel-spectrogram, which is resilient to regular signal transformation techniques. In this way, the compression or transformation of audio cannot accurately restore the initial clear speech. Furthermore, WaveGuard reduces the audio quality of the synthetic speech. As a result, VSMask maintains a high level of robustness against denoising and signal transformation approaches. For sophisticated attackers, it is possible to recover the raw audio by reverse engineering if they have full knowledge about VSMask. Other approaches, for example, randomized smoothing [26], adversarial training [27] and diffusion model [28] are potential solutions to evade VSMask.

6 DISCUSSION AND LIMITATIONS

Defense in Real-world Scenarios. Fig. 9(a) displays the setup of physical world VSMask protection. The speaker is reading transcripts from VCTK dataset in a quiet room. Meanwhile, VSMask is deployed on a laptop 50 cm away from the speaker. At the same time, an eavesdropper is stealthily recording the speech audio through an iPhone 13 one meter away from the speaker and VSMask. Then, the eavesdropper will synthesize the speaker's voice in the recorded speech by AdaIN-VC model. Meanwhile, we put a sound level meter close to the microphone to measure the volume of sounds. The background noise in the room is *40* dBA, and we maintain the speech loudness around *75* dBA. While the victim is talking in quiet room, the attacker can eavesdrop clear speech signal and generate qualified deepfake voice.

Notably, in real-world scenarios, different from cyberspace, the perturbation δ is usually distorted during over-the-air transmission. To overcome the challenge, we implement a band pass filter (500 ∼ 4000 Hz) to filter vulnerable frequencies, and room impulse response (RIR) filter to mitigate the signal distortion in airborne transmission [29]. When VSMask is applied in a new environment, the RIR filter should be redesigned. Fig. 9(b) shows the protection impact under real-world scenarios. When the average perturbation loudness of VSMask reaches 46 dBA, it can effectively compromise the target voice synthesis model without affecting the speech intelligibility. In comparison, random white noise with the same volume fails to degrade the similarity score much. When the perturbation
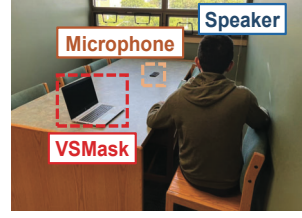
(a) Physical world experiment setup.
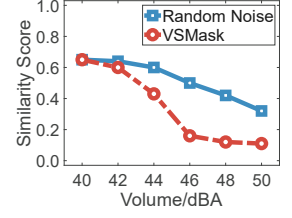
(b) VSMask protection performance in real-world scenario.

Figure 9: We implement VSMask in real-world scenario, where it shows much stronger protection impact than white noise.

volume reaches 50 dBA, VSMask can successfully reduce the ASR to 0% while most synthetic speech generated from white noise masked speech can still bypass the ASV model. The experiment is launched in a quiet environment. If the environment is noisy (Volume > 50 dBA), the synthesized speech quality will also be affected even without VSMask. In particular, adaptive attackers may use a microphone array to denoise VSMask perturbation by phase cancellation [30]. We can enhance the spatial complexity and increase the difficulty of VSMask perturbation cancellation by utilizing multi-channel loudspeakers.

Real-time Feasibility. Training a predictive model for one speaker on an NVIDIA A6000 GPU costs less than 5 minutes. In the inference process, one single computation of the predictive model costs ∼ 13 milliseconds on a common CPU. We also measure the time delay caused by bidirectional transformation between audio signal and mel-spectrogram, which is ∼ 200 milliseconds. Therefore, the total time cost for generating voice perturbation is around 0.2 seconds, and the latency of 0.4 seconds is sufficient for VSMask to achieve real-time applicability.

Adversarial Training. Adversarial training can effectively improve the robustness of deep learning models [27]. Adaptive attackers can apply adversarial training to improve the similarity between the synthetic voice and the victim speaker protected by VSMask. However, adversarial training inevitably reduce the synthesis performance when dealing with clear speech audios. Moreover, recent work [31] shows that adversarial training takes up to 10 days with a powerful GPU group even for a mini dataset. Therefore, given VSMask's real-time protection feature, adversarial training is not a realistic solution that can be adopted by attackers to defeat VSMask.

Limitations. VSMask has several limitations. First, specially designed adaptive attacks are not considered in this paper, e.g., adversaries who have full knowledge about VSMask. The adaptive attackers can purify the speech by a well-designed neural network or diffusion models, and we leave the exploration of these evasive approaches in future work. Second, we only launch real-world evaluation in a single room in the physical experiments. In different environments, the performance of VSMask might be different. Third, VSMask cannot provide 100% successful protection regardless of the voice synthesis model. A few synthetic speech can still pass ASV systems under pure black-box setup. Finally, the proposed method does not provide a formal privacy guarantee for human speech. This paper aims to conduct empirical privacy analysis with VSMask and provide a practical solution to enhance speech privacy in real-time physical and digital scenarios.

7 RELATED WORK

7.1 Voice Synthesis

For traditional VC methods, it is necessary to train the model with parallel speech data from the source speaker and victim speaker. Chen et al. [32] propose a voice conversion model based on Gaussian Mixture Models (GMM). Erro et al. [33] apply a weighted frequency wrapping method to change in voice in the speech. But these methods suffer from over-smoothing [34] and unnatural voice. It is also difficult to collect the parallel speech data from different speakers [35, 36].

To address such challenges, Variational Autoencoder (VAE) is introduced to convert human voice in the speech [37]. Compared with traditional voice conversion models, VAE is able to learn independent speech content regardless of the speaker identity. Therefore, VAE shows better conversion efficiency and requires no parallel speech data for training. Hsu et al. [38] propose a voice conversion model that generates converted voice frame-by-frame. AdaIN-VC [16] significantly improves the converted speech quality by adding an extra adversarial training stage. Cycle-GAN [39] and Star-GAN [40] also apply adversarial learning method in VAE based voice conversion. One of the state-of-the-art work is AutoVC [17], which is a voice conversion model that introduces bottleneck mechanism into VAE. It successfully achieves zero-shot conversion and balances the speaker disentanglement and voice conversion quality via well-designed bottleneck. Moreover, generative adversarial network (GAN) itself can be applied to clone human voice [41]. But it achieves limited conversion quality. TTS can also synthesize speech with real human voice. Wavenet [42] introduces a generative model to output raw audio according to the input text. Tacotron 2 [43] further inverts Wavenet generated spectrograms with attention. But Tacotron 2 can only generate speech with one default speaker voice. Deep Voice 2 [44] is a speech generative model that supports multiple speaker, while Deep Voice 3 [45] uses a encoder-decoder based neural network. It can apply TTS synthesis with over 2,400 different voices from LibriSpeech. SV2TTS [18] introduces a TTS model that enables speech synthesis for an arbitrary speaker. It shares similar speaker encoder module as voice conversion models. Wenger et al. [6] comprehensively evaluate the voice synthesis threat in the real world scenarios.

7.2 Adversarial Speech Attacks and Defenses

Existing work demonstrates that neural networks can be easily fooled by adversarial examples. Carlini et al. [46] point out that adversarial speech samples can compromise ASR systems. Schönherr et al. [13] demonstrate that ASR can be compromised by perturbation out of human hearing range. Adversarial speech samples can be hidden in speech [47] or music [48], and deliberately crafted to deceive speech recognition systems. In some cases, these adversarial samples can be made imperceptible to the human ear [49], posing security threats in both digital and physical environments.

Similarly, speaker recognition models are also vulnerable to adversarial attacks [50, 51]. Chen et al. [52] proposes FAKEBOB, an adversarial attack successfully compromise speaker recognition systems under black-box and over-the-air attack scenarios. Zhang et al. [53] also successfully spoof commercial speaker verification systems by well-designed adversarial speech examples.

Offline adversarial speech attacks only work when the entire speech is available. To launch real-time attacks toward streaming speech signal, Lu et al. [14] introduce universal adversarial perturbation, which could force the ASR models to output the target transcript regardless of the real input is. AdvPulse [54] successfully fools ASR system by a short adversarial "pulse". The adversarial perturbations successfully mislead the ASR system without signal synchronization. But these universal perturbations work only if the crafted output is short. Gong et al. [55] introduce a real-time attack targeting streaming voice signal. This real-time attacks works well for keyword detection applications with a pre-defined dictionary. SpecPatch [56] proposes an adversarial attack could mute the upcoming voice commands and compromise long speech samples. VoiceCamo [11] proposes a real-time attack based on a predictive model. It leverages the past speech signal as input and predict the most effective perturbation to generate adversarial example along with the upcoming speech.

7.3 Speech Privacy Enhancement

Qian et al. [57] leverage voice conversion to hide speaker's identity. McAdams [58] and Vaidya et al. [59] mask human speech by signal processing to enhance speech privacy. Han et al. [60] propose a differential privacy based fast speech de-identification system to provide formal privacy guarantee on an untrusted server. Emotionless [61] is a privacy-preserving speech analysis tool for voice assistants based on differential privacy. In contrast, this paper aims to develop a practical speech protection system to offer privacy enhancement for real-world speech applications.

Meanwhile, adversarial examples can be leveraged to enhance speech privacy. V-Cloak [62] presents a timbre-preserving model to achieve voice anonymization, but its protection performance against voice synthesis attacks is not evaluated. Attack-VC [10] introduces a defense method based on adversarial examples, which could protect offline human speech against voice synthesis attacks. VSMask introduces a novel combination of universal perturbation and predictive models to protect the real-time speech from voice synthesis attacks.

8 CONCLUSION

Voice synthesis attack is a major threat to the voice controllable systems. This research aims to prevent the attackers from synthesizing the protected voices. Specifically, we explore the feasibility of real-time defense against voice synthesis. By implementing real-time predictive model and universal perturbation header, VSMask successfully protects real-time speech signal in both online and offline real-world scenarios. Moreover, VSMask mitigates the audio distortion in protected speech via the weight-based amplitude constrained mechanism. We evaluate VSMask's voice protection performance under different real-world scenarios, and the experimental results show that VSMask can effectively defend against voice synthesis attacks in live speech scenarios.

REFERENCES

[1] B. U. Mahmud and A. Sharmin, "Deep insights of deepfake technology: A review," *arXiv preprint arXiv:2105.00192*, 2021.

[2] "Amazon's alexa impersonates grandma, creeping out internet," https://www.bloomberg.com/news/articles/2022-06-23/amazon-s-alexa-impersonates-grandma-creeping-out-the-internet, June 23, 2022.

[3] "Voice guidance & navigation," https://www.readspeaker.ai/applications/voice-guidance-and-navigation/, Accessed on August 15, 2022.

[4] "Voiceprint: The new wechat password," https://blog.wechat.com/2015/05/21/voiceprint-the-new-wechat-password/, May 21, 2015.

[5] Y. Wang, H. Guo, and Q. Yan, "Ghosttalk: Interactive attack on smartphone voice system through power line," in *Network and Distributed Systems Security (NDSS) Symposium*, 2022.

[6] E. Wenger, M. Bronckers, C. Cianfarani, J. Cryan, A. Sha, H. Zheng, and B. Y. Zhao, "" 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*, 2021, pp. 235–251.

[7] M. E. Ahmed, I.-Y. Kwak, J. H. Huh, I. Kim, T. Oh, and H. Kim, "Void: A fast and light voice liveness detection system," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 2685–2702.

[8] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[9] X. Li, X. Wu, H. Lu, X. Liu, and H. Meng, "Channel-wise gated res2net: Towards robust detection of synthetic speech attacks," *arXiv preprint arXiv:2107.08803*, 2021.

[10] C.-y. Huang, Y. Y. Lin, H.-y. Lee, and L.-s. Lee, "Defending your voice: Adversarial attack on voice conversion," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 552–559.

[11] M. Chiquier, C. Mao, and C. Vondrick, "Real-time neural voice camouflage," in *International Conference on Learning Representations*, 2021.

[12] "71+ instagram statistics 2022," https://https://www.demandsage.com/instagram-statistics/, November, 2022.

[13] L. Schönherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," *arXiv preprint arXiv:1808.05665*, 2018.

[14] Z. Lu, W. Han, Y. Zhang, and L. Cao, "Exploring targeted universal adversarial perturbations to end-to-end asr models," *arXiv preprint arXiv:2104.02757*, 2021.

[15] "Equal loudness contours (iso 226-2003)," https://chart-studio.plotly.com/~mrlyule/16/equal-loudness-contours-iso-226-2003/#/, Accessed on August 10, 2022.

[16] J.-c. Chou, C.-c. Yeh, and H.-y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," *arXiv preprint arXiv:1904.05742*, 2019.

[17] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.

[18] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in neural information processing systems*, vol. 31, 2018.

[19] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[21] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[22] K. Murano, S. Unagami, and F. Amano, "Echo cancellation and applications," *IEEE Communications Magazine*, vol. 28, no. 1, pp. 49–55, 1990.

[23] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[24] J. Zhuang, T. Tang, Y. Ding, S. C. Tatikonda, N. Dvornek, X. Papademetris, and J. Duncan, "Adabelief optimizer: Adapting stepsizes by the belief in observed gradients," *Advances in neural information processing systems*, vol. 33, pp. 18 795–18 806, 2020.

[25] S. Hussain, P. Neekhara, S. Dubnov, J. McAuley, and F. Koushanfar, "WaveGuard: Understanding and mitigating audio adversarial examples," in *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, Aug. 2021, pp. 2273–2290. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/hussain

[26] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *international conference on machine learning*. PMLR, 2019, pp. 1310–1320.

[27] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[28] S. Wu, J. Wang, W. Ping, W. Nie, and C. Xiao, "Defending against adversarial audio via diffusion model," *arXiv preprint arXiv:2303.01507*, 2023.

[29] H. Yakura and J. Sakuma, "Robust audio adversarial example for a physical attack," *arXiv preprint arXiv:1810.11793*, 2018.

[30] J. Meyer and C. Sydow, "Noise cancelling for microphone arrays," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1997, pp. 211–213 vol.1.

[31] B. Zheng, P. Jiang, Q. Wang, Q. Li, C. Shen, C. Wang, Y. Ge, Q. Teng, and S. Zhang, "Black-box adversarial attacks on commercial speech platforms with minimal information," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 86–107.

[32] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, "Voice conversion with smoothed gmm and map adaptation," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[33] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2009.

[34] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[35] S. H. Mohammadi and A. Kain, "Voice conversion using deep neural networks with speaker-independent pre-training," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 19–23.

[36] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify the modulation spectrum in hmm-based speech synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 290–294.

[37] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[38] C. C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," 2017.

[39] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv preprint arXiv:1711.11293*, 2017.

[40] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[41] Y. Gao, R. Singh, and B. Raj, "Voice impersonation using generative adversarial networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2506–2510.

[42] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[43] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[44] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," *Advances in neural information processing systems*, vol. 30, 2017.

[45] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017.

[46] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE security and privacy workshops (SPW)*. IEEE, 2018, pp. 1–7.

[47] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, "Hidden voice commands," in *25th USENIX Security Symposium (USENIX Security 16)*. Austin, TX: USENIX Association, Aug. 2016, pp. 513–530. [Online]. Available: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/carlini

[48] X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter, "{CommanderSong}: A systematic approach for practical adversarial voice recognition," in *27th USENIX security symposium (USENIX security 18)*, 2018, pp. 49–64.

[49] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *International conference on machine learning*. PMLR, 2019, pp. 5231–5240.

[50] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, "Fooling end-to-end speaker verification with adversarial examples," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 1962–1966.

[51] Y. Gong and C. Poellabauer, "Crafting adversarial examples for speech paralinguistics applications," *arXiv preprint arXiv:1711.03280*, 2017.

[52] G. Chen, S. Chenb, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real bob? adversarial attacks on speaker recognition systems," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 694–711.

[53] L. Zhang, Y. Meng, J. Yu, C. Xiang, B. Falk, and H. Zhu, "Voiceprint mimicry attack towards speaker verification system in smart home," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 377–386.

[54] Z. Li, Y. Wu, J. Liu, Y. Chen, and B. Yuan, "Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020, pp. 1121–1134.

[55] Y. Gong, B. Li, C. Poellabauer, and Y. Shi, "Real-time adversarial attacks," *arXiv preprint arXiv:1905.13399*, 2019.

[56] H. Guo, Y. Wang, N. Ivanov, L. Xiao, and Q. Yan, "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*, 2022, pp. 1353–1366.

[57] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X.-Y. Li, "Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity," in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, 2018, pp. 82–94.

[58] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the mcadams coefficient," *arXiv preprint arXiv:2011.01130*, 2020.

[59] T. Vaidya and M. Sherr, "You talk too much: Limiting privacy exposure via voice input," in *2019 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2019, pp. 84–91.

[60] Y. Han, Y. Cao, S. Li, Q. Ma, and M. Yoshikawa, "Voice-indistinguishability – protecting voiceprint with differential privacy under an untrusted server," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 2125–2127. [Online]. Available: https://doi.org/10.1145/3372297.3420025

[61] R. Aloufi, H. Haddadi, and D. Boyle, "Emotionless: Privacy-preserving speech analysis for voice assistants," *arXiv preprint arXiv:1908.03632*, 2019.

[62] J. Deng, F. Teng, Y. Chen, X. Chen, Z. Wang, and W. Xu, "V-cloak: Intelligibility-, naturalness-& timbre-preserving real-time voice anonymization," *arXiv preprint arXiv:2210.15140*, 2022.

APPENDIX

A PREDICTIVE MODEL

According to VSMask system design in Section 4, the input and output of the predictive model are supposed to be 2-D vectors with the same dimension and different lengths. Inspired by the model architecture in VoiceCamo [11], we apply a similar down-sampling and up-sampling architecture in the VSMask predictive model. To ensure the best prediction performance, we test different kernel sizes, strides, and layer numbers. We start with kernel size (2, 2) and determine that kernel size (3, 3) achieves the best trade-off between the protection performance and model complexity. Then, we carefully adjust the strides and layer numbers to correct the final output vector dimension. Finally, we use a tanh activation layer to normalize the predictive perturbation. The detailed parameter configuration is listed in Table 4. The architectural parameters vary according to the different preprocessing approaches in different voice synthesis models.

B IMPACT OF WEIGHTS ON AUDIO QUALITY

We carefully select the weight in Problem (5) for different frequency bands to ensure they have the same protection performance. In this section, we investigate how the weight-based amplitude constraint mitigates the noise perceptibility in protected audios.

In human study, we also ask the volunteers to compare the clear speech and protected speech without weight constraint. Fig. 10 shows the perceptibility difference between protected audio with

Table 4: The detailed network parameters of VSMask real-time predictive model when targeting AdaIN-VC model.

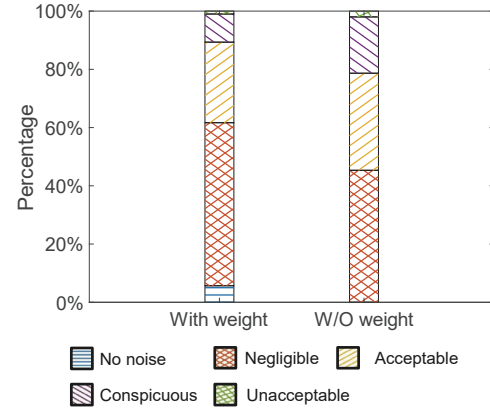| Type	Input Size	Kernel Size	Stride
	1×512×100	3×3	(1,2)
Down-sampling	32×512×50 128×512×25 256×256×13 256×128×7 512×64×4 512×32×2	3×3	(2,2)
Up-sampling	512×16×1 256×32×2 128×64×4 64×128×8 32×256×16	3×3	(2,2)
tanh	1×512×32	N/A	N/A

Figure 10: Human perceptibility for VSMask protected speech with and without weight-based constraint.

and without weighted constraint. When we implement weight-based constraint, the perturbation in medium frequency is weakened. In this way, human ears are less sensitive to the noise in the speech samples. Over 60% answers consider the noise is negligible or imperceptible, which means that VSMask perturbation does not affect the utterance and voice in the speech at all. There are also 30% answers present they can tolerate the noise in the speech despite they are conscious of the perturbation.

In comparison, if we remove the weight mechanism, it becomes easier to sense the perturbation. Less than 50% answers demonstrate they can neglect the perturbation in the speech, while more than 20% of results indicate that there is conspicuous noise in the audio. A few responses point out that the noise is too strong which affects the utterances and voice in the speech. Therefore, the human study results demonstrate that VSMask's weighted constraint can significantly reduce the audio distortion in the protected audios.