Inaudible Adversarial Perturbation: Manipulating the Recognition of User Speech in Real Time

Xinfeng Li
Zhejiang University
xinfengli@zju.edu.cn

Chen Yan†
Zhejiang University
yanchen@zju.edu.cn

Xuancun Lu
Zhejiang University
xuancun_lu@zju.edu.cn

Zihan Zeng
Zhejiang University
zengzh@zju.edu.cn

Xiaoyu Ji†
Zhejiang University
xji@zju.edu.cn

Wenyuan Xu
Zhejiang University
wyxu@zju.edu.cn

***Abstract*—Automatic speech recognition (ASR) systems have been shown to be vulnerable to adversarial examples (AEs). Recent success all assumes that users will not notice or disrupt the attack process despite the existence of music/noise-like sounds and spontaneous responses from voice assistants. Nonetheless, in practical user-present scenarios, user awareness may nullify existing attack attempts that launch unexpected sounds or ASR usage. In this paper, we seek to bridge the gap in existing research and extend the attack to user-present scenarios. We propose VRIFLE, an inaudible adversarial perturbation (IAP) attack via ultrasound delivery that can manipulate ASRs as a user speaks. The inherent differences between audible sounds and ultrasounds make IAP delivery face unprecedented challenges such as distortion, noise, and instability. In this regard, we design a novel ultrasonic transformation model to enhance the crafted perturbation to be physically effective and even survive long-distance delivery. We further enable VRIFLE's robustness by adopting a series of augmentation on user and real-world variations during the generation process. In this way, VRIFLE features an effective real-time manipulation of the ASR output from different distances and under any speech of users, with an *alter-and-mute* strategy that suppresses the impact of user disruption. Our extensive experiments in both digital and physical worlds verify VRIFLE's effectiveness under various configurations, robustness against six kinds of defenses, and universality in a targeted manner. We also show that VRIFLE can be delivered with a portable attack device and even everyday-life loudspeakers.**

I. Introduction

Automatic speech recognition (ASR) enables computers to transcribe human speech and is essential in a wide range of voice applications such as voice assistants (VAs) and audio transcription APIs [1], [2]. Prior studies have shown that ASR models are vulnerable to adversarial examples (AEs) that sound benign to humans but are recognized incorrectly by models. As stealthiness is a basic requirement for AEs, existing works largely focus on reducing the audibility of AEs so that they might not cause human suspicion when being heard [3], [4]. In addition, the class of inaudible attacks [5], [6] avoids being perceived by human ears using high-frequency ultrasound/laser. However, few of them have considered attacks in user-present scenarios, where users may notice unexpected events of the ASR service and can mitigate the attack's consequence. For instance, though AEs and inaudible attacks may not sound suspicious, a voice assistant will always provide feedback (e.g., vocal prompt or LED blinking) after receiving voice commands. Alert users may still notice the false wake-up or abnormal feedback caused by an attack and speak remedy commands to correct the mistake, limiting the attack's impact in real life.

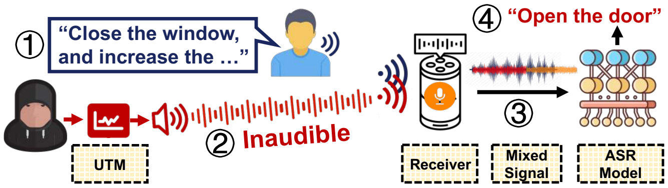


Fig. 1: ①When a user uses the ASR service, ②an adversary injects inaudible adversarial perturbations crafted based on the ultrasonic transformation model (UTM) into a receiver. ③The mixed signal of the user command (blue) and demodulated perturbations (red & yellow) can ④fool the ASR model into the adversary-desired intent.

In this paper, we aim to propose VRIFLE[1]—an *inaudible adversarial perturbation* (IAP) attack that can extend to this scenario. Its basic idea is to inject IAPs while the user speaks to the ASR service and alter the recognition result in real time, as shown in Fig. 1. Since the voice assistant itself is responding to user commands (e.g., LED blinking), tampering with the user's speech is less noticeable at this time. But such a user-present scenario also imposes higher requirements on attack stealthiness because users are more sensitive to environmental sounds while using ASR services. Moreover, given that adversaries have no prior knowledge of user's speech content and timing, this critical scenario necessitates that VRIFLE exhibits a high level of universality to guarantee the achievement of the adversary-desired intent in any context. Therefore, we envision VRIFLE as a truly inaudible and robust framework for real-time IAPs delivery, which can also address variable user factors,

†Chen Yan and Xiaoyu Ji are the corresponding authors

[1] Demo: https://sites.google.com/view/Vrifle

Network and Distributed System Security (NDSS) Symposium 2024
26 February - 1 March 2024, San Diego, CA, USA
ISBN 1-891562-93-2
https://dx.doi.org/10.14722/ndss.2024.23030
www.ndss-symposium.org

such as speech content, vocalization time, speech volume, and environmental conditions, while remaining physically effective even at long distances or using portable/everyday-life devices. Overall, materializing VRIFLE that attains the above goals is challenging in three aspects.

- *How to achieve adversarial perturbations that are* ***universal*** *while* ***completely inaudible*** *to* ***user auditory****?*

The trade-off between universality and stealthiness has been a long-standing challenge in audio AE attacks. Almost all previous works have prioritized stealthiness and introduced imperceptibility constraints during optimization, such as ϵ and L2-Norm [7], or by adjusting audio forms, e.g., designing it as short pulses [8]. Nonetheless, this greatly compromises the universality of adversarial perturbations and they are still audible and can be heard when users are nearby. We seek to implement an inaudible adversarial perturbation beyond the human auditory range (20 Hz$\sim$20 kHz) in an ultrasound-based attack manner [5], which can make microphones receive our IAP by exploiting their inherent nonlinearity vulnerability. As such, IAPs are no longer limited by stealthiness constraints, holding a vast optimization space with more feasible solutions. Unlike the audible-band perturbations devised to be short to mitigate user auditory, our IAPs enable the adversary to significantly increase their length, which further expands the optimization scope and facilities highly universal attacks.

- *How to alter the recognition of user speech in real time despite the presence of* ***user disruption****?*

Although we have bypassed *user auditory*, realizing such an attack against ASR in real time faces a few more challenges. *User disruption* cannot be ignored in this scenario, which includes: ❶ The user's speech can disrupt the intent of IAPs when both audio signals are superimposed. While universal AEs [8], [9] are shown to resist this case, our preliminary investigation validates that direct ultrasound-based attacks will fail due to such interference. ❷ User commands can be much longer (e.g., 5s) than 0.5s audible-band short perturbations that affect only a few input frames, thus the exceeded user instructions will impact the entire ASR transcription. ❸ Users may notice that malicious behavior being executed and therefore block the attack by issuing remedy commands. In addition, there are user-induced factors that make *user disruption* more complex and can compromise IAPs' effectiveness, including unpredictable content and timing of user speech, as well as the influence of the user's environment and speaking habits on speech reverberation and loudness.

To address these issues, we augment the optimization process of IAPs by using multiple speech clips in public corpus, introducing randomness within the preset time range, as well as considering the various user's speech loudness and reverberation. Thereby, VRIFLE can be applied in a content-agnostic, synchronization-aided, user factors-robust manner. Moreover, we overcome *user disruption* by materializing both silence and universal perturbations in the targeted manner to ensure the arbitrary utterance length cannot pose impacts on adversary-desired intent, without requiring any knowledge. Based on the above design, adversaries can present two more hidden attack strategies, involving *No-feedback Attack* and *Man-in-the-middle Attack* in the threat model.

TABLE I: Compared with existing works

Method	**Constraint**‡	**Auditory**$^{\natural}$	**Disruption***	**Dist.**†
Carlini. [10]	–	Noise	✗	1.5m
Abdullah [11]	–	Noise	✗	0.3m
CW [7]	L2-norm, ϵ	Speech	✗	✗
Schönherr [3]	Psyc.	Speech	✗	✗
Comman. [12]	ϵ	Song	✗	1.5m
Qin. [4]	Psyc.	Speech	✗	–
Meta-Qual [13]	L2-norm, ϵ	Song	✗	4m
FakeBob [14]	ϵ	Speech	✗	2m
AdvPulse [9]	L2-norm, ϵ	Ambient	◗	2.7m
SpecPatch [8]	L2-norm	Pulse	◗	1m
Ours	**None**	**Inaudible**	●	**10m**

(i) ‡: The constraints used to guarantee imperceptibility during optimization. "–" means the method only considers incomprehensibility to humans. ϵ means limiting the absolute magnitude of perturbations with a constant ϵ. L_2-Norm means adding an L_2-Norm term in the objective function. "Psyc." means psychoacoustic hiding. "None" means no stealthiness constraints. (ii) ♮: The objective ***user auditory*** of AEs. Ambient means ambient sounds. (iii) ⋆: ●: fully tackles ***user disruption***. ◗: tackles case ❶. ✗: fails by ***user disruption***. (iv) †: ✗: the attack is not physically available. –: not reported.

- *How to guarantee inaudible adversarial perturbations are* ***physically effective*** *after ultrasonic delivery?*

Though inaudible attacks have demonstrated voice command injection using ultrasound and laser [6], *it is unknown whether fine-grained IAPs can be delivered via such signals* as the ultrasound channel is reported to be lossy and distorted [15]. Thus, maintaining the effectiveness of IAPs after undergoing a series of modulation, transmission, and demodulation processes in the physical world is not trivial based on prior AEs [16]. Ultrasound is intrinsically distinct from audible sounds in the high-directional propagation and varying soundfield. Additionally, the nonlinear distortion, anomalous noises, and hardware-induced instability that are unique to ultrasound make existing acoustic channel modeling methods inapplicable.

To overcome the challenge, we make the first attempt to establish an ultrasonic transformation model, which consists of tackling variable ultrasound-induced anomalous noises, obtaining ultrasound frequency response (UFR), and enabling location-variable attacks. Based on this transformation, we can precisely estimate VRIFLE's pattern of ultrasonic delivery during its optimization, thereby making it physically effective and survive long-distance delivery. Moreover, to enable more covert IAP attacks with portable devices and off-the-shelf loudspeakers, we implement a narrow bandwidth upper-sideband modulation (USB-AM) mechanism to ensure the attack range and inaudibility of VRIFLE with simplified devices.

Tab. I compares VRIFLE with several existing works. We conduct extensive experiments in both digital and physical worlds to evaluate VRIFLE's effectiveness under various configurations (e.g., extend attack range to 10m) and robustness against six kinds of defenses. Our single silence IAP muting up to 27,531 unseen user utterances, likewise, universal IAP altering 18,956, proving VRIFLE's universality. Our design also expands the attack methodology to more covert portable attack devices and everyday-life loudspeakers, enabling the VRIFLE delivery in a stealthier form. Our contribution can be summarized as follows:

- To the best of our knowledge, VRIFLE is the first universal inaudible adversarial perturbation attack that can extend to scenarios when users use ASR services, revealing a new attack surface against ASR models. VRIFLE is completely inaudible, holds vast optimization space, and enables long-range attacks (10m).
- We make the first attempt to establish an ultrasound transformation model, which overcomes the unique challenges in the ultrasound channel and precisely characterizes it, enabling our fine-grained IAPs delivery to be physically effective.
- We conduct extensive experiments under various configurations in the digital and physical world to validate the effectiveness, robustness, and universality of VRIFLE, and validate the attack using portable/everyday-life devices.

II. Background

A. Automatic Speech Recognition

Automatic speech recognition (ASR) systems, e.g., voice assistants, receive and recognize speech commands; then perform execution according to certain rules. Hidden Markov models (HMM) [17] and dynamic time warping (DTW) [18] are two traditional statistical techniques for performing speech recognition. With the development of deep learning, the end-to-end neural ASR models have gone mainstream, such as RNN-T [19] and DeepSpeech2 [20]. A typical end-to-end ASR system pipeline includes four main components: ① *Spectrum generator*: converts raw audio into spectrum features, e.g., Filter Bank (Fbank), Mel-Frequency Cepstral Coefficients (MFCC), etc. ② *Neural acoustic model*: takes spectrums as input and outputs a matrix of probabilities over linguistic units (e.g., phoneme, syllable, or word) over time. For instance, English ASR is widely modeled with 29 basic units (also known as tokens), including characters a˜z, space, apostrophe, and blank symbol ϕ. ③ *Decoder*: generates possible sentences from the probability matrix, also optionally coupled with an n-gram language model to impose word-level constraints. The Connectionist Temporal Classification (CTC) module is a typical decoder that sums over all possible alignments that may reduce to the same token sequence, whereby "oϕkkϕaϕy" and "oϕkϕaaϕyy" are regarded as the same "okay". ④ *Punctuation and capitalization model*: formats the generated text for easier human consumption.

B. Audio Adversarial Examples

Adversarial examples (AEs) [3], [7], [12], [13] use specialised inputs created with the purpose of confusing a neural network, resulting in the misclassification of a given input. In the audio domain, by adding a crafted perturbation δ with some constraints ϵ throughout the original benign audio x, the ASR model will be fooled to transcribe a perturbed speech into the targeted text y_t, e.g., "take the picture". To craft an adversarial example, an adversary may leverage the optimization function:

$$\begin{aligned} &minimize\ \mathcal{L}(f(x+\delta), y_t) + \alpha \cdot \|\delta\|_p \\ &s.t.\ \delta \in [-\epsilon, \epsilon]^n, (\epsilon{<}0.01) \end{aligned} \tag{1}$$

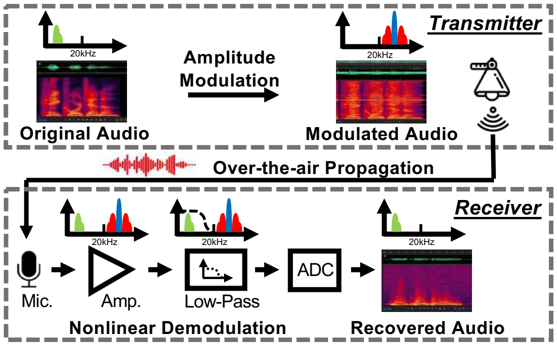


Fig. 2: Diagram of inaudible attacks (carrier: blue, baseband: green).

where the ASR functions as $f(\cdot)$ that takes an input waveform and outputs the probability matrix. $\mathcal{L}(f(\cdot), y_t)$ is the CTC loss function denoting the distance between the model output of the adversarial example and the target. $\|\cdot\|_p$ means the L_p norm. α is a penalty term to limit the L_p. ϵ denotes the upper bound of the perturbation. Recently, the concept of universal adversarial perturbation is proposed, making AEs valid regardless of the user commands. To make the AEs more concealed, the creating approaches are extended to psychoacoustic hiding [3], [4] and shorter pulses [9]. However, existing efforts cannot fundamentally avoid being perceived by human ears.

C. Ultrasound-based Attacks

Inaudible attacks modulate the audio baseband on high-frequency carriers to the inaudible band of human ears (>20 kHz) and exploit microphones' nonlinear vulnerability, so that ASRs can receive the malicious audio while humans cannot perceive it. Recently, inaudible attacks have been extended from ultrasonic carrier [5], [21] to various forms, such as solid conduction [22], laser [6], capacitor [23], power line [24], etc., forming a class of highly threatening and comprehensive covert attacks. We take the representative ultrasound-based attack [5] to present the principle of inaudible attacks shown in Fig. 2. First, the original audio is double-sideband (DSB) modulated on an ultrasound carrier via amplitude modulation (AM). Second, the DSB-AM audio is emitted from the ultrasonic transducer and propagates over the air. Third, after the microphone receives the signal, audio modulated on the high-frequency carrier will be recovered into the audible band *before the low-pass filters and ADC* due to nonlinear effects of the microphone's diaphragm and amplifier. Thus, though the ultrasound carrier is finally filtered, the demodulated audio still survives and functions to ASR. The nonlinear demodulation is formulated as follows:

$$S_{out}(t) = \sum_{i=1}^{\infty} k_i s^i(t) = k_1 s(t) + k_2 s^2(t) + k_3 s^3(t) + \ldots \tag{2}$$

where $s(t)$ and $S_{out}(t)$ indicate the input AM signal and amplifier's output, respectively. The even order terms, i.e., k_2, k_4 are the key in recovering the original audio [25]. Notably, such an ultrasound channel is lossy as the recovered audio samples differ from original ones. Our investigation demonstrates that the channel is also challenging to model III-D.

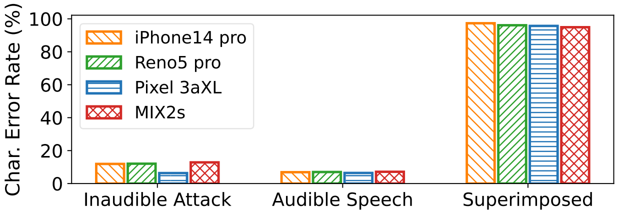


Fig. 3: CER of four recording devices under three settings.

D. Threat Model

Attack Scenarios: We consider attacks in user-present scenarios where the user may notice unintended events of ASR services. Such scenarios involve two entities:

Victim: The victim user is alert to any strange sounds (e.g., noise, music, pulses) within human auditory. The user can speak an arbitrary command to the smart speaker. Once the user notices attacks, he/she can speak a remedy command to the smart speaker.

Adversary: The adversary prepares IAPs for specific intentions offline and alters user command in real time by delivering IAPs ①at a distance from the victim with an ultrasonic transmitter through the window, ②physically close with a handheld portable device, or ③with a preset off-the-shelf loudspeaker. The adversary's goals are providing wrong information to intelligent voice customer service, compromising VAs to execute malicious commands or be in denial-of-service mode, etc. The adversary can attack more covertly with two strategies: *1) No-feedback Attack*: Prevent the user from hearing VA's vocal prompt by "Mute volume and turn off the WiFi". *2) Man-in-the-middle Attack*: Once the user's intent is satisfied, the attack may be much less suspicious, i.e., while delivering the adversarial perturbation and alter user commands, adversaries can record the user commands and then replay it by traditional ultrasound-based attack means.

Attacker Capability: Distinct from the previous works [3], [4], [7], [12]–[14], [26] that require the user's speech samples in advance to craft adversarial perturbation, we assume the adversaries have no knowledge of what the user will speak during performing attacks. In line with the widely adopted settings in prior works [4], [7], [8], [12], [13], we assume the attackers have prior knowledge of the target ASR model for obtaining the gradient information during optimization. The adversaries have access to the user's recording device, e.g., borrow a smartphone of the same brand, based on which the adversary can model the ultrasonic transformation, and then create the IAP in advance. We assume adversaries have the flexibility to deploy the hidden ultrasonic transmitter nearly or at a distance, and the recording device is in its line of sight. Additionally, adversaries can also utilize stealthy portable devices and off-the-shelf loudspeakers in everyday-life scenarios to deliver VRIFLE.

III. PRELIMINARY INVESTIGATION

A. Failure of Traditional Inaudible Attacks

Given the purpose of avoiding alerting users, directly injecting malicious commands into ASR systems using laser- [6] or ultrasound-based [5], [21] inaudible attacks is intuitive. Although laser-based attacks can reach an 100m attack range, we choose ultrasound instead of laser for three practical reasons: (1) The laser spot on the microphone is visible and will alert users immediately; (2) The laser-based attack requires strict line-of-sight alignment; and (3) The severe channel distortion of laser-delivered attacks may nullify fine-grained adversarial perturbations.

To examine whether traditional ultrasound-based attacks can manipulate ASRs into recognizing the modulated malicious commands while users are speaking, we need to ensure that the ultrasonic carrier frequency is optimal. Therefore, we first employ an ultrasonic Vifa [27] to launch a wide-range carrier sweeping from 20∼40 kHz. By analyzing the signal-to-noise ratio (SNR) of demodulated basebands, we justify the optimal frequency of four recording devices, i.e., iPhone14 pro: 24.7 kHz, Reno5 pro: 27.7 kHz, Pixel 3aXL: 25.6 kHz, and MIX2s: 25.1 kHz, respectively. This result is consistent with DolphinAttack [5], which reveals most devices' optimal attack carrier frequency is around 25 kHz (22.6∼27.9 kHz). In this way, we set the default carrier frequency to 25 kHz, whose advantages are two-fold: (1) Due to lower airborne attenuation, 25 kHz also benefits longer-range attacks than high-frequency carriers (e.g., 40kHz); (2) Moreover, 25 kHz as one of the most typical parameters for commercial ultrasonic transducers that cost as low as 0.14$ per unit [28], making the attack cost-effective.

Although the optimal attack frequency is determined, traditional ultrasound-based attacks still fail due to *user disruption*. Specifically, we select 10 text-to-speech commands listed in Tab. IX (e.g., "turn on airplane mode") as the basebands. Four smartphones 50 cm away serve as recording devices that recover the AM signal into audible-band speech. For benign command samples, we randomly select 20 utterances from the popular fluent speech commands dataset [29] to be played via a loudspeaker and recorded by identical smartphones. We also perform simultaneous emissions of both signals, so they are superimposed on each other. For each recording device, we collected $10 \times 20 = 200$ mixed samples and calculated each sample's character error rate (CER) through the Azure speech-to-text API [1]. As shown in Fig. 3, the direct ultrasound-based attacks and benign audio are well recognized by ASR models, with average CER of 10.8% and 6.88%, respectively. Nevertheless, once attack emission and user's voice coincide, the attack performance (i.e., 10 malicious commands as the target transcription) will severely degrade to an average CER up to 96.01%, even if we have boosted its power[2]. We believe it is a consequence that when ASRs process the mixed samples, each sampling point of the malicious signal sequence is affected by the human voice, making the acoustic features extracted by the ASR deviate from adversaries' anticipation.

B. Ultrasonic Adversarial Perturbation Delivery

We envision that the above failure can be addressed by leveraging the vulnerability of ASR models to craft universal adversarial perturbations. Notably, it is promising to deliver the perturbation in an ultrasound-based manner to eventually reach the goal, i.e., the adversary can alter any user commands into a targeted one while guaranteeing entirely inaudible to

[2] To facilitate ultrasound-based attacks, we set the volume up to 95 dB, and that of audible benign speech is 70∼75 dB.

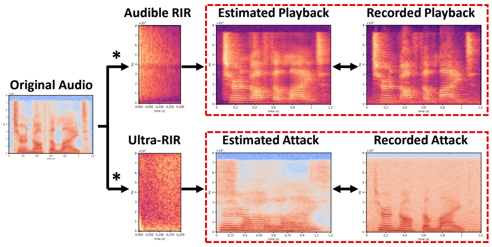


Fig. 4: Comparison between the audible RIR and ultra-RIR in estimating the digital-to-physical transformation.

the victim. However, we find the well-trained perturbations that are effective in digital domain all fail after being directly modulated and emitted by the ultrasound-based attack method (results are also given in §V-C1, *G2*).

Since the ultrasonic channel is lossy and distorted, to obtain a perturbation that can still effectively tamper with user commands after a series of processing based on ultrasound-based attack mechanisms and over-the-air delivery, i.e., the pipeline shown in Fig. 2, we need to precisely model the transformation from a perturbation in the digital domain to its physical version. However, generalizing an AE from the digital to the physical world is inherently difficult, which has been proved by substantial research in both computer vision and audio community [7]–[9], [13], [16], [30], [31]. This issue in the audio domain refers to the fact that played-out speech samples are subject to signal distortion and environment interference (i.e., reverberation, attenuation, and noises). Previous audible-band works [16], [32] have paid efforts in simulating the physical world by adopting room impulse response (RIR) during the AE optimization process to close the gap between the digital and physical world. Moreover, no work has yet been proposed on modeling ultrasonic delivery. We are motivated to investigate the feasibility of applying audible-band modeling technologies to our unique ultrasonic case.

C. Attempts at Ultrasound Delivery Modeling

In this subsection, we elaborate on applying two potentially feasible modeling methods for our case.

1) Modeled by room impulse response: Inspired by the success of audible-band AEs [16], [32] drawing on the ability of RIR, which describes the reverberation and attenuation during audible sound propagation, we envisage that a similar RIR idea can generalize to characterize the ultrasound transformation process. Specifically, they exploit existing RIR databases [33], [34] by convolving random RIR clips with digital adversarial signals in the optimization process, simulating the audio recorded by the receiver in various scenes, e.g., large concert hall and narrow corridor. Therefore, we modulate the ideal impulse signal as the baseband on an ultrasound carrier and receive it on the recording device. With the obtained "ultra-RIR", we perform convolution with the original audio, whose output are expected to well represent the actually demodulated inaudible attack's result. As a comparison, we also conduct similar operations via a JBL loudspeaker for audible audio. Fig. 4 shows that the estimated audible audio with RIR is very close to the actual playback. However, for the inaudible aspect, there is significant gap between the recorded attack audio and the estimated using ultra-RIR. We believe the reason for such a mismatch is that the RIR rationale relies on the linear time-invariant (LTI) system prerequisite. However, the transformation is nonlinear because ultrasound-based attacks leverage microphones' nonlinearity vulnerability.

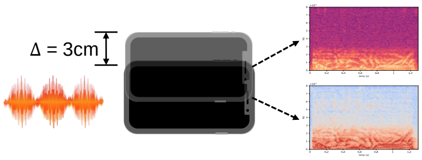


Fig. 5: Illustration of displacement-induced changes in recorded audio.

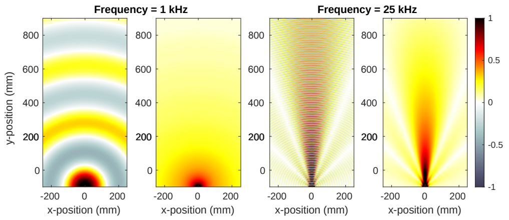


Fig. 6: 2-D sound fields simulation for comparing the audible wave with ultrasound.

2) Modeled by Neural Network: Since RIR is originally designed for LTI systems, neural networks with excellent nonlinear fitting capability should work well given their success in various tasks, e.g., speech denosing [35] and image printing distortion [30]. Considering that an adversary expects a practical transformation model with minimal effort (i.e., dataset requirements) while guaranteeing its generality, we implement a multi-layer perception model (MLP) with only 60k parameters, using 120-second aligned original and ultrasound-based attack audio pairs. We find that the MLP can achieve a generalized capability of mapping digital-to-physical world spectrums between unseen pairs, but with position-dependent constraints. As shown in Fig. 5, a slight position displacement (3 cm) leads to an apparent change (i.e. bringing anomalous noises) in the recovered baseband, which can cause the trained network to fail to estimate the recorded audio at various positions. Overall, although MLP builds a functional mapping for the nonlinear ultrasound transformation in a fixed relative position, it is too restricted due to the nature of ultrasound (cf. §III-D). Besides, adopting distance d and angle θ as conditional network parameters might help, but collecting data for each position is endless.

D. Challenges in Ultrasonic Delivery Modeling

The above attempts' failure drives us to look into the root cause of why modeling ultrasonic transformation is challenging. Ultrasound is intrinsically distinct from audible sounds due to its much higher frequency, and ultrasonic delivery leverages microphones' nonlinearity. We also summarize the following characteristics:

- *Ultrasound-induced Noise:* The ultrasound carrier continuously forces the diaphragm to vibrate, probably resulting in anomalous noise in recorded attack alike Fig. 4&5. Combined with such variable ultrasound fields, a slight displacement (e.g., 3 cm) can lead to different audio patterns.

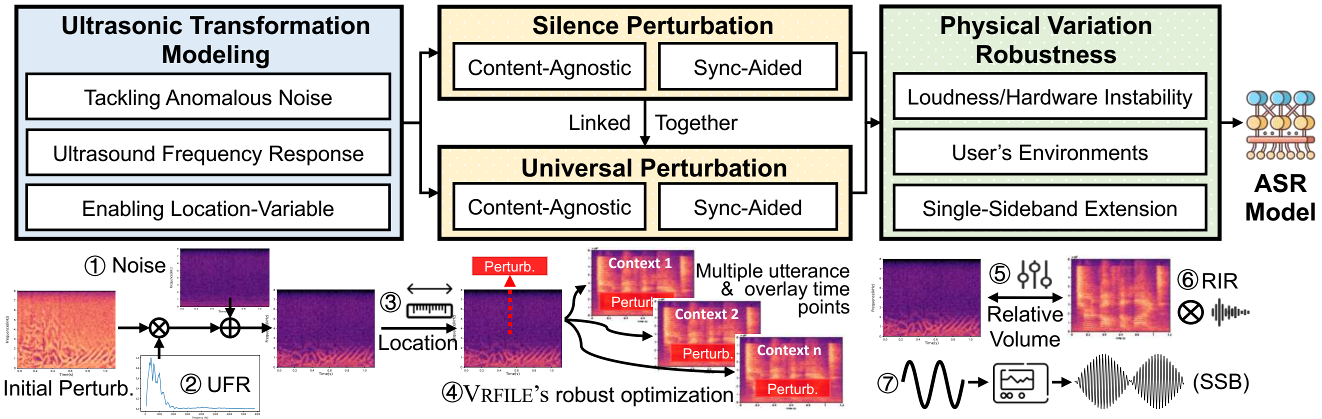


Fig. 7: Workflow of VRIFLE. ①-③: the ultrasonic transformation precisely describes the perturbation changes during physical delivery. ④: the transformed perturbation is involved into optimization for silence and universal attack purpose. ⑤-⑦: we boost the attack's physical-world robustness from multiple aspects.

- *Nonlinear Distortion:* Eq. 2 indicates the I/O relationship of nonlinear demodulation, in which the factors k_i is unknown and varies with recording devices [15].
- *Varying Soundfield:* As shown in Fig. 6, ultrasound field (25 kHz) is significantly more directional and changes more dramatically than audible waves (1 kHz) due to the much shorter wavelength.
- *Hardware-induced Instability:* Ultrasound-based attacks rely on a series of signal processing and sophisticated devices, thus bringing instability due to hardware imperfection.

IV. Design of VRIFLE

A. Overview

Design Goal. To manipulate ASRs while being used by users, adversaries shall create universal IAPs. However, they face the following challenges to obtain and deliver such perturbations:

Ultrasound Complexity (C1). Modeling the ultrasonic delivery is unprecedented compared to the audible-band RIR mimics, because ultrasound fundamentally differs from audible sound as listed in §III-D, including (i) ultrasound-induced anomalous noises, (ii) nonlinear distortion, (iii) varying sound field, and (iv) hardware-induced instability.

User-ASR Connection (C2). ASR systems always respond to the user after receiving a command. Adversaries need to suppress the impact of *user disruption*, i.e., break down the user-ASR connection by IAPs that can silence user's excessively long speech and remedy commands.

User Variation (C3). Since adversaries cannot exactly know the user speech's content, timing, or length, naively mixed speech signals will lead to undesirable ASR transcriptions. The tailored IAP needs to be universal while facing arbitrary user commands and superimposed time points.

Physical Robustness (C4). The adversary also faces several factors that are variable in physical attacks, such as user loudness, hardware instability, and the user's environment (i.e., with different reverberations). We also extend the modulation method for reducing unexpected sound leakage.

To achieve adversaries' goal while addressing the aforementioned challenges, we propose VRIFLE with unique technical design. This design includes: (1) tackling ultrasound complexity to deliver physically effective IAPs and therefore addressing *user auditory* (cf. §IV-B); (2) overcoming *user disruption* to achieve real-time manipulation of ASR (cf. §IV-C, IV-D); (3) boosting attack stealthiness and practicality (cf. §IV-E). The optimization workflow of VRIFLE is exhibited in Fig. 7.

Problem Formalization. Unlike the audible-band AE attacks subject to stealthiness constraints, we achieve inaudible perturbations delivery using ultrasound modulation. Thus, we avoid the narrow constraints in Eq. 1, e.g. $\epsilon < 0.01$, where the IAP's optimization space can reach the maximum upper bound: $\delta \in [-1, 1]^n$. We believe a broad optimization space possesses more feasible solutions, facilitating a universal attack. Combined with our core objective: fooling ASRs to recognize the superimposed speech of user voice and perturbation $x + \delta$ as the adversary-desired transcription y_t. This basic idea can be optimized via the following formulation:

$$\begin{aligned} &minimize\ \mathcal{L}(f(x+\delta), y_t) \\ &s.t.\ \delta \in [-1,1]^n\ and\ x+\delta \in [-1,1]^n \end{aligned} \tag{3}$$

B. Ultrasonic Transformation Modeling

As shown in Fig. 8, our modeling exploits the ⑤additive property of the baseband audio m's nonlinear transformation $H(f)m(f)$ and the ultrasound-induced anomalous noise n, and then ⑥yields estimated audio $\hat{m} = H(f)m(f) + n$ that is highly similar to the actual recorded audio $\widetilde{m}$. In this subsection, we elaborate on our divide-and-conquer strategy of implementing ultrasonic transformation modeling that overcomes problems (i)-(iii) corresponding to *ultrasound complexity (C1)*. Based on this, we can deliver physically effective IAPs via the steps ①~③ in Fig. 7. We address the problem (iv) in §IV-E.

1) Tackling Anomalous Noises: Ultrasound-based attacks modulate the baseband m with $s(t) = A[1 + m(t)]c(t)$, where regardless of the energy of m, the carrier signal $c(t) = cos\omega_c t$ always emits and forces the microphone diaphragm into vibration, appearing abnormal noises [15]. Our experiment also demonstrates that although the recorded s varies with m, the

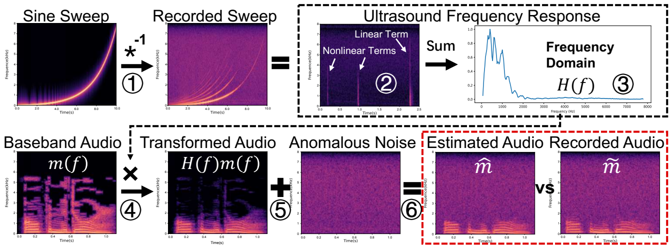


Fig. 8: Ultrasonic transformation modeling. **1st row:** Procedure to obtain the ultrasonic frequency response (UFR). **2nd row:** High similarity between the estimated audio and actually recorded audio (the red box) proves its effectiveness.

anomalous noise pattern is almost decided by the carrier. The nature of the ultrasound field further results in noise variation with different injection angles θ and distances d, showing irregular patterns. Therefore, due to such variation, neural networks fail to learn a stable mapping of the digital-to-physical domain. We denote the noise $n(\theta, d) = f_n(\theta, d, s)$, where f_n is the projection of the ultrasound signal s to the recorded abnormal noise n at different positions. In practice, an attacker can sample the variable anomalous noises by simply emitting the ultrasonic carrier. We collect a lightweight noise dataset using 25 kHz ultrasound (without modulation) of 1m at varying angles, forming a set U_n of 25 pieces of 10-second noises.

2) Ultrasonic Frequency Response: Recalling the reasons for LTI system-based RIR's failure in §III-C1 of modeling unprecedented ultrasonic delivery, except for anomalous noises, the inability to describe the nonlinear demodulation process is also a key factor. The adversaries aim to achieve robust and adaptive attacks with minimal effort, i.e., building an efficient transformation that can well estimate the demodulated pattern of a given digital perturbation after ultrasonic delivery in Fig. 8 (red box). Fig. 8 also depicts the recorded audio derived after inaudible signal injection, whose energy is clearly concentrated in the low-frequency band compared to the original audio [21]. Although nonlinearity exists, we are driven to obtain an ultrasonic frequency response (UFR) that characterizes the inaudible acoustic energy conversion at different frequencies.

We first overcome ultrasound-induced noises that hinder us from obtaining an accurate frequency response by adopting the sine sweep technique [36], which can ignore components uncorrelated to the sweep signal during processing. We use it to generate a fast 10s sweep ranging from 50$\sim$7800 Hz, which is carefully chosen for diminishing hardware imperfection, and record it on the receivers, shown in Fig. 8①. Thus, we can obtain the UFR $H(f)$ by deconvolution ($*^{-1}$). Notably, as shown in Fig. 8②, it does shield the effects of noises and focuses on the frequency response measurement, which decouples the linear and nonlinear terms. We sum these terms up in Fig. 8③, forming a holistic frequency-domain UFR of the received perturbations δ as $\overline{\Delta}(f) = H(f)\Delta(f)$, where $\Delta(f) = \mathcal{F}(\delta(t))$; $\mathcal{F}$ means Fourier Transform.

3) Enabling Location-Variable: Uneven ultrasound field makes MLP-based method in §III-C2 difficult to estimate transformation from arbitrary-position attack. As for efficient UFR, we believe that combining it with ultrasound $s(d, t)$ propagation process [37] will empower to render more adaptive attacks:

$$H(f, d) = H(f) \cdot e^{-a_0 \omega_c{}^n d}, \; n \in [1, 2] \tag{4}$$

where a_0 is a medium-dependent attenuation parameter, ω_c is the carrier's frequency. Moreover, the energy variation caused by different injection angles is hard to model under such a changing sound field. We overcome this issue by conducting sine sweeps at different angles θ similar to §IV-B1 and get 25 pieces of 10-second sweep clips. Consequently, the collection of a complete set of UFRs and anomalous noises for subsequent optimization requires approximately 8.3 minutes. Overall, with a pair of UFR $H_\theta(f, d)$ and noise clip $n(\theta, d)$ from the same location, we can well estimate the digital perturbation into its recording. However, to obtain a location-variable perturbation, we shall modify the expression of Eq. 3 and find the perturbation via robust training:

$$\underset{\delta}{argmin} \underset{h_\theta \sim U_H, n \sim U_n}{\mathbb{E}} [\mathcal{L}(f(x + h_\theta(d) * \delta + n), y_t] \tag{5}$$

where we use time-domain expression $h_\theta(d) * \delta$ to indicate the transformed perturbation's waveform, as $H_\theta(f, d)\Delta(f) = \mathcal{F}[h_\theta(t, d) * \delta(t)]$ obeys the time convolution theorem. We randomly select the UFR $H_\theta(f, d)$ and noise n pairs from U_H and U_n during the optimization process to mimic actually delivering the inaudible adversarial perturbation at different locations. As we fully take ultrasound's inaudibility advantages, the experiment results also validate the optimization space is large enough to craft a robust perturbation effective under varying UFRs and noises.

C. Silence Perturbation

Given the failures of prior works faced with *user disruption*, specifically ❷ the challenge of excessively long instructions, and ❸ the potential counteraction through remedy commands, we believe that the solution lies in silencing the user instructions, i.e., breaking down the user-ASR connection when necessary. Based on our ultrasonic transformation modeling, adversaries can materialize physically effective silence perturbations. These perturbations can alter arbitrary user instructions to blank (“ ”) in a targeted manner, effectively rendering the ASR system unresponsive to the user instructions. We observe that implementing silence perturbations offers several advantages, including: (1) When altering long user commands with short target intent, such as “start recording” (case ❷). The silence perturbation can be linked alongside the universal perturbation in an *alter-and-mute* fashion (cf. §IV-D) so that the ASR will output adversary-desired transcription; (2) it guarantees that users cannot meddle in running malicious operations by issuing remedy commands (case ❸), even if they notice the presence of attacks; (3) it can render ASR services in the denial-of-service condition, preventing users from using them normally.

Fig. 9(b) depicts the diagram of a robust silence perturbation ξ, which is expected to superimpose over any benign content like Fig. 9(a) and leads the final transcription of ASR to blank y_b (“ ”). The length of ξ is empirically set to 5s based on our experiments, for which we balance the duration of common speech instructions and the optimization overhead. For the case of excessively long user utterances, we address them in the generation process by repeating the perturbation.

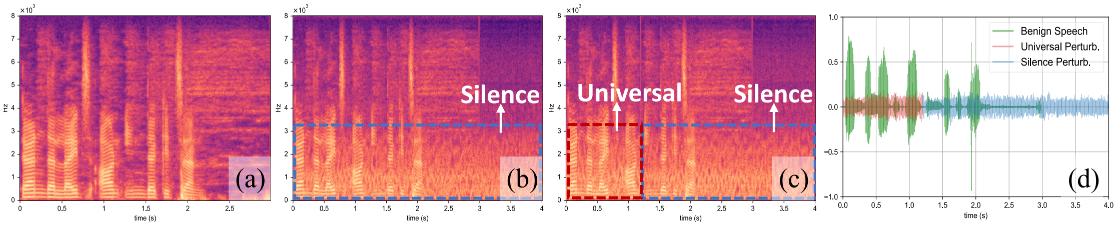


Fig. 9: Diagram of VRIFLE attacking a benign speech (a) with the silence (b) and universal goals in an *alter-and-mute* manner (c&d).

To craft such a content-agnostic ξ, we improve the penalty-based expectation function to find the silence perturbation over a group of common voice commands U_x, as shown in Fig. 7④.

$$\underset{\xi}{argmin} \underset{h_\theta \sim U_H, n \sim U_n, x \sim U_x}{\mathbb{E}} [\mathcal{L}(f(\mathcal{S}_x + h_\theta(d) * \xi + n), y_b)] \quad (6)$$

where $\mathcal{S}_{(\cdot)}$ means randomly shifting the user utterances x for introducing randomness to the superimposed time within a preset T:100 ms. U_x is elaborated in the experimental setup §V-A2. It is more practical than the case where an AE and user speech are required to be perfectly aligned. The details of content-agnostic and synchronization are given in §IV-D.

D. Universal Perturbation

Different from the proof of concept of universal AEs against a CNN-based speech command classification model presented in [9] by exploiting the temporal insensitivity of CNNs, the RNN-based models widely deployed on commercial ASR services are more difficult to attack. This difficulty arises because end-to-end ASRs, such as DeepSpeech [20], employ connectionist temporal classification (CTC) that calculates the loss between a continuous speech feature sequence and a target transcription, making it context-dependent. Consequently, when introducing subtle perturbations in different contexts, it is often difficult to ensure that the CTC losses of multiple mixed signals will simultaneously converge to the desired target.

1) Content-Agnostic: We believe that the reasons why previous audible-band AEs struggle to tamper with large amount of speech content are two-fold: *user auditory* and *user disruption*. To avoid being noticed by users, prior adversarial perturbations are limited by imperceptibility constraints and signal forms (e.g., with short length and subtle amplitude). Consequently, these perturbations are fragile and easily defensible. In contrast, our IAP delivery is completely inaudible via ultrasound modulation. Thus, the perturbation's length and amplitude are unconstrained, maximizing its optimization space. We fully use the advantages to generate a universal perturbation that can alter substantial short utterances into adversary-desired intent, e.g., a 1.2s δ tailored for "open the door".

However, for excessively long speech or possible subsequent remedy commands in user-present scenarios (*user disruption* ❷-❸), the adversary should resort to the silence perturbation in §IV-C, which can cooperate well with the universal perturbation in an *alter-and-mute* manner. As depicted in Fig. 9(c) and (d), when the universal perturbation δ is combined with a well-trained silence perturbation $\hat{\xi}$, the former can apply to alter the user commands, and the latter will mute the subsequent user commands or remedies. As illustrated in Fig 7④, we determine the optimal δ by optimizing the following expectation function:

$$\underset{\delta}{argmin} \underset{h_\theta \sim U_H, n \sim U_n, x \sim U_x}{\mathbb{E}} [\mathcal{L}(f(x + h_\theta(d) * \overline{\delta : \hat{\xi}} + n), y_t)] \quad (7)$$

where $\overline{\delta : \xi}$ means the universal perturbation δ followed by a crafted silence perturbation $\hat{\xi}$. U_x is the same subset used for generating silence perturbations, whose details are given in §V-A2.

2) Synchronization-Aided: Although the universal perturbation can deceive the ASR with any victim's speech into adversary-desired intent, an adversary can hardly deliver attacks synchronously when the victim vocalizes. Out of attack practicality, we propose a VAD-based synchronization mechanism to achieve real-time manipulation, which avoids continuous AE broadcasting or assuming an adversary always ready for attacking. Specifically, we employ a microphone to record the user's voice. Once detecting the user's speech via voice activity detection (VAD), our program automatically triggers the emission of the prepared perturbation. Based on our experiments, the delay is impacted by three stages of our real-time pipeline: (1) from user vocalizing to being detected by the running VAD program (5∼20 ms); (2) software-to-hardware IAP triggering (5∼15 ms); (3) ultrasound propagation (0∼30 ms). Due to the delay uncertainty, we consider bringing the time randomness of a range T into our optimization, whose upper bound is empirically set to 100 ms. For a direct reference, the average overall delay when attacking at 4m is around 27 ms, far below the maximum tolerable delay (100 ms) preset during optimization. Particularly, the recording of user speech can also be utilized to present a more covert attack by inaudibly replaying user-desired commands, as "*Man-in-the-middle Attack*" stated in §II-D. By integrating the above-mentioned optimization objectives, we further craft the universal IAP through the below expectation:

$$\underset{\delta}{argmin} \underset{h_\theta \sim U_H, n \sim U_n, x \sim U_x}{\mathbb{E}} [\mathcal{L}(f(x + \mathcal{S}_{h_\theta(d) * \overline{\delta : \hat{\xi}} + n}), y_t)] \quad (8)$$

where $\mathcal{S}_{(\cdot)}$ mimics VRIFLE can be superimposed on victim speech at random time points (Fig. 7④) within the preset T.

E. Physical Robustness

1) Loudness Adaptive and Hardware Instability: When conducting physical attacks, VRIFLE is able to handle the challenges of ultrasound nature based on our digital-to-physical transformation in §IV-B. However, the loudness of the victim's speech varies with context or emotion, and hardware instability still exists. These will result in difficulty maintaining our inaudible perturbation in effectively altering the victim's voice if the mutual energy relationship between the two is inconsistent with the optimization process. As shown in Fig. 7⑤, we introduce relative volume augmentation into the crafting process, which exploits a hyper-parameter β denoting a range of user speech's volume and thereby brings randomness to the mutual relationship between user voice and perturbation.

2) Attack at Different Environments: Although ultrasound-based attacks directly inject into recording devices' microphones and are reverberation-free, the audible-band human voice still goes through multi-path reflections and ambient noises in different environments. To alter user commands regardless of the impact of scenes, we apply random RIR and noise clips from the Aachen Impulse Response (AIR) Database [33] in Fig. 7⑥, including small, medium, large rooms and corridors for user speech augmentation.

3) Single-Sideband Extension: Although VRIFLE can achieve real-time manipulation of the ASR output very covertly using sophisticated devices (e.g., narrowband ultrasonic transducers and signal generators) at long distances through windows or doors, we aim to accomplish highly stealthy attacks even in close proximity to the victim by utilizing everyday-life loudspeakers or portable attack devices. However, the simple amplifiers, sound cards, and off-the-shelf loudspeakers exhibit poor suppression of intermodulation and harmonics of high-frequency DSB-AM signals. Namely, they present increased nonlinearity, resulting in sound leakage (cf. §VII). To enable attacks with portable devices and loudspeakers (cf. §V-D), we adopt single-sideband amplitude modulation (SSB-AM), which removes one of the sidebands based on the Hilbert transform [38]. Compared to DSB-AM, SSB-AM has only half the bandwidth, rendering higher transmission efficiency. Importantly, it mitigates the intermodulation between different sideband frequencies, making the sound less prone to leakage than DSB-AM at the same energy level. Specifically, we employ upper sideband modulation (USB-AM), formalized as $S(t) = mcos\omega_c t - \hat{m}sin\omega_c t + cos\omega_c t$, rather than lower sideband modulation (LSB-AM), as the former exhibits better inaudibility in our experiments and more details are given in Appendix §A.

Overall, the algorithm of VRIFLE is described in Algorithm 1, Appendix §D, where we demonstrate the optimization process of crafting VRIFLE from scratch.

V. EVALUATION

A. Experiment Setup

1) Overview: We implement VRIFLE using PyTorch [39] on a Ubuntu 20.04 server with Intel Xeon 6226R 2.90GHz CPU and NVIDIA 3090 GPU. Based on our experiments, we empirically set the default configuration as $\delta = 1.2s$, $\xi = 5s$, $\epsilon = 1$, $0.5 \leq \beta \leq 1.5$, sync range T=100 ms, $maxEpoch$=800. Adam optimizer [40] is used to speed up our convergence. For evaluating VRIFLE's effectiveness in fooling ASR while users use it, we select the end-to-end DeepSpeech2 [20] as the target model and conduct experiments in both digital and physical scenarios.

2) Dataset: We adopt the typical Fluent Speech Command Dataset [29] to examine the effectiveness of VRIFLE, including 30,046 voice command samples. We randomly selected 896 samples from the 10-person validation set given in the dataset, with each speaker contributing around 90 utterances on average. These samples are used to craft our perturbation. The remaining unseen 29,150 samples are used to evaluate VRIFLE under various settings.

3) Hardware: We employ a signal generator (SIGLENT SDG6032X) [41] to modulate the created IAPs, a power amplifier (NF HSA4015) [42] to enable long-range delivery, and a custom ultrasound transducer array to emit the modulated IAPs. The recording devices to be tested include Google Pixel 3aXL, iPhone14 pro, MI Mix2s, OPPO Reno5 pro, and ReSpeaker Mic array v2.0 [43], where all model versions are released in the last five years. Moreover, we evaluate attacks with a self-made portable device and a loudspeaker in §V-D.

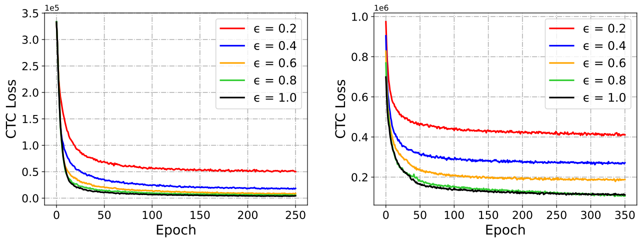
4) Metrics: (1) We use the success rate (SR) to indicate the percentage of VRIFLE successfully altering user commands and matching target transcriptions in all attempts. (2) We use character error rate (CER), a representative metric in ASR tasks, to indicate the adversary's ability to tamper with user commands from the character level; a lower CER represents a more effective attack. (3) Signal-to-Noise Ratio (SNR) and $L2$-distortion are vital for audible-band AEs because of the imperceptibility requirements. SNR: the ratio of benign audio power to the perturbation power. $L2$: the sum of squared amplitude. AEs with a low SNR and high $L2$ are more likely to be noticed, and vice versa.

B. Digital Attack Performance

As our attack focuses on real-world scenarios, where physical disturbances always exist, we incorporate the effects of physical conditions by employing our ultrasonic transformation modeling to guarantee that digital attack performance has physical significance.

1) Impact of Optimization Space: Since the delivery of VRIFLE is inaudible, it facilitates the unconstrained advantage of setting ϵ up to 1 (i.e., the normalized audio's upper bound) for universal attacks. We further explore attack capability under different ϵ upper bounds, both universal and silence IAPs. We optimize silence perturbations according to $\epsilon = 0.2, 0.4, 0.6, 0.8, 1.0$, respectively, aiming to tamper the user instructions to the blank. In addition, we obtain the universal perturbations expected to alter user commands to "open the door" with the same settings. CTC loss convergence curves are shown in Fig. 10. We observe that the crafting process can converge faster as the ϵ (i.e., the optimization space) increases in both tasks. After ϵ reaches 0.8, the convergence rate approaches the maximum. Then we estimate the physical delivery of both perturbations via an unseen transformation model (i.e., a pair of UFR and anomalous noise not involved in training). The transformed perturbations are further superimposed on every test voice command sample. Results listed in Tab. II show that, in addition to the faster convergence, a larger ϵ significantly boosts the universality of VRIFLE. It can successfully alter 18,946 samples into "open the door" and mute 27,531 user commands into blank " ", which highlights VRIFLE features a highly universal capability.

2) Comparison of Convergence Overhead and Audibility Cost for the Universality Goal: The unconstrained advantage of VRIFLE ($\epsilon = 1$) empowers its high universality. We further compare it with 3 classical audible-band AEs (i.e., CW [7], Qin [4], and SpecPatch [8]) regarding the cost for achieving



(a) Silence Perturbation (b) Universal Perturbation

Fig. 10: CTC loss curves of silence and universal perturbations during the optimization process under varying ϵ.

TABLE II: The number of successfully silenced/altered test speech samples under different ϵ upper bounds

Upper Bound (ϵ)	0.2	0.4	0.6	0.8	1.0
Silence Perturb.	1,591	8,095	17,064	24,832	27,531
Universal Perturb.	649	5,268	13,085	16,726	18,946

the same universality goal of each method. We reproduce these works strictly following their instructions. We set *two goals*: creating a single perturbation that can alter 1) one or 2) five commands based on each method. Notably, for audible-band AEs, we employ RIRs for physical simulation to be consistent with our default setup. We specify the minimal upper bounds ϵ of CW, Qin, and SpecPatch to $0.03, 0.05, 0.05$, respectively, based on which the three methods can maintain universality for 5 commands, i.e., finally converge to the target transcript "Open the door". We examine also the CTC loss convergence speed of 4 methods. The normalized loss curves in Fig. 11 clearly show that VRIFLE (in red) converges within the fewest iterations among 4 methods; SpecPatch (in blue) converges slowest as it is devised to be short (0.5s). Specifically, we list the overall duration for each methods to final convergence for altering 5 commands—VRIFLE: 1.63 min, CW: 6.52 min, Qin: 9.16 min, and SpecPatch: 35.38 min. VRIFLE converges faster because (1) we reduce optimization complexity by only picking 5 random UFR/noise pairs rather than all ultrasonic channel data per iteration; (2) VRIFLE can quickly find feasible solutions due to its broad optimization space. In addition, Tab. III demonstrates the SNRs and $L2$-distortion values of audible-band AEs under different universality goals. All SNRs of these AEs are low due to a compromise of physical robustness and imperceptibility, with the highest SNR down to 22 dB. Moreover, if the goal number increases, audible-band AEs are bound to get louder and more easily heard.

3) Different Target Commands: Given that adversaries may launch attacks for various purposes, they will craft different adversarial perturbations accordingly. In this experiment, we first train 10 universal perturbations referring to typical malicious commands [44] listed in Appendix §C Tab. IX along with the silence perturbation. Then we apply VRIFLE to 7,200 benign commands to validate its effectiveness, amounting to 72,000 samples. We count the success rate when transcription outputs match the target commands correctly. In addition, we also count CERs over all samples. We find no significant performance varying with target transcripts, where most targets derive a 100% SR and 0% CER (7 out of 10). The lowest SR is still up to 92.82%, corresponding to "Mute volume and turn off the WiFi". Moreover, it is worth noting that the highest CER of these targets is still down to 0.50%, suggesting VRIFLE can tamper with user commands well from the character level. Due to page limitations, the details are listed in Appendix C Tab. IX.

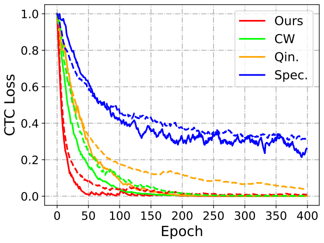


Fig. 11: CTC loss curves. Compare the convergence speed of VRIFLE with 3 classical audible-band AEs. Dashed lines: train a perturbation that can simultaneously alter 5 voice commands; likewise, Solid lines: alter 1 command, thus converging faster than the former.

TABLE III: Audibility cost

Method	**Goal**	**SNR**	$L2$
CW	1 5	22dB 18dB	1.31 2.24
Qin	1 5	22dB 17dB	1.42 2.67
Spec.	1 5	8.5dB 8.4dB	122 123

C. Physical Attack Performance

We perform extensive physical experiments to evaluate the practical performance of VRIFLE under different conditions, i.e., w/o our modeling, distances, environments, recording devices, etc. In the physical experiments, we set the target intent as "open the door", the attack distance 4m away from the recording devices with the injection angle pointing to their bottom microphones as the default configuration unless otherwise specified. Except for the experiments about different scenes, the rest are conducted in a laboratory of approximately 13.6m×5.2m with slight HVAC noises. We employ a custom ultrasonic transmitter for inaudible adversarial perturbation delivery. A loudspeaker plays the audible benign speech samples, and the ambient noise level is around 38 dB. We also deploy a VAD-based program in conjunction with a microphone connected to the laptop to trigger IAPs delivery using the synchronization-aided design. This ensures real-time triggering when audible benign speech initiates. Our real-world attack scenario is given in Appendix §B, Fig. 19.

1) Ablation Experiments w/o Transformation Modeling: To validate the effectiveness of our ultrasonic transformation modeling, we apply 3 strategies to craft IAPs. In addition, we apply direct ultrasound-based attacks as the baseline group (*G1*). The first strategy is an optimization without transformation, i.e., $h_\theta(d)*\delta+n$ in Eq. 8 is degraded to simple δ during the crafting process (*G2*). Similarly, the second strategy uses a low-pass filter, reducing the precise transformation to a filter that allows signal components below 3 kHz to pass (*G3*). The third strategy is crafting a perturbation with *our transformation*, i.e., VRIFLE (*G4*). We carry out experiments with synchronization-aided emission of both benign audio and attacks. We select 40 benign utterances to be played via loudspeaker, and finally collect 480 mixed samples (120 per group) by repeating the operation three times for minimizing errors. Tab. IV lists each group's success rate and average CER, which remarkably denotes that our modeling can well describe the digital-to-physical transformation during optimization. Approximating the transformation as a low-pass filter can also generate a physically available perturbation with 21.67% SR and 19.39%

TABLE IV: Ablation of w/o transformation modeling

Metrics	Baseline (*G1*)	Without (*G2*)	Low-pass (*G3*)	With (*G4*)
SR	0% (0/120)	0% (0/120)	21.67% (26/120)	100% (120/120)
CER	95.7%	78.93%	19.39%	0%

TABLE V: Different attack scenes

Scene	Office	Lounge	Laboratory	Corridor
SR	100% (40/40)	95% (38/40)	100% (40/40)	92.5% (37/40)
CER	0%	0.79%	0%	1.04%

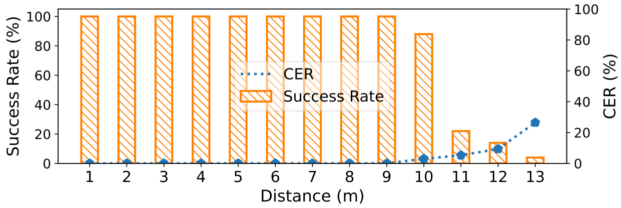


Fig. 12: VRIFLE's performance at different distances.

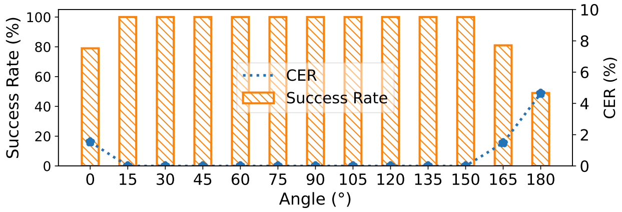


Fig. 13: VRIFLE's performance at different angles.

CER. The attack performance decreases to 0% in *G1* and *G2*. However, attacks without modeling still outperform the baseline (*G1*) from the CER perspective due to leveraging the model vulnerability.

2) Different Attack Distances: Attacks in the audible band are constrained by concealment, resulting in perturbations that cannot be delivered with more extensive ranges. By contrast, our attack delivery via ultrasound modulation can apply substantial power, which overcomes the attenuation nature of ultrasound. We adjust the amplifier gain so that the high-frequency beam's energy reaching microphones is maintained, thus ensuring the effectiveness of VRIFLE. Specifically, we conduct experiments at the ultrasonic transmitter away from the receiving device within 1m∼13m (1m interval), where the maximum power at 10m∼13m is approximately 3.2 Watt. We randomly select 40 voice commands and play them at each location. We repeat the perturbation superimposed on the benign command 3 times and totally collect 1,560 samples, with 120 per distance, respectively, as well as feed them into the ASR model. We count the success rate and CER in Fig. 12, where VRIFLE is very effective within 1m∼9m as the SRs are up to 100% and CERs are down to 0%. The SR is 88.7% and CER is still down to 3.25% at 10m. Besides, we observe the attack performance decrease at 11m∼13m. We believe this is due to the ultrasound attenuation, which makes the perturbation less significant to the ASR model. We also discuss this issue in §VII.

3) Different Attack Angles: In this experiment, we keep the recording device's bottom microphone spatially within the ultrasound beam's coverage and set the attack distance to 2.5m. We rotate the recording device from 0°to 180°at 15°intervals, among which 90°means the ultrasound directly points to the bottom microphone. Under each angle, we play 40 benign commands and emit the universal IAP. Eventually, we collected 520 mixed audio signals from 13 angles. As shown in Fig. 13, although ultrasound is highly directional, we find that there is no significant difference with 100% success rate among different angles within 15°∼150°. As the deployed location of bottom microphones varies with different phones, therefore attack performance is not symmetrical with angles (i.e., 79% at 0°and 49% at 180°). Overall, as most voice-interface devices nowadays are equipped with omnidirectional microphones, VRIFLE can be effective as long as the beam can cover the bottom microphone.

4) Different Scenes: To examine the effectiveness of VRIFLE in different environments, our experiments include a small office (2.4m×2.6m, 36 dB), medium lounge (6.3m×3.8m, 42 dB), large laboratory (13m×5.2m, 38 dB), and narrow corridor (60m×2m, 44 dB). In these scenes, the reverberation pattern of audible sound varies with space size. Our configuration consists of a transmitter-to-device distance of 4m and a loudspeaker-to-device distance of 1m, which mimics the standard user interaction distance, except the distance of 2.5m for the small office due to its limited size. We also play 40 audible benign samples and superimpose VRIFLE on them for once. Then we collected 160 samples from 4 spaces. As shown in Tab. V, we find no significant difference between these scenarios, as our design considers such physical variation.

5) Different Ambient Noises: We perform ambient noise-related experiments in our laboratory, where noises of 4 typical scenes are involved, i.e., cafeteria (people chatting), office (keyboard typing), lab (machine running), and outdoor (wind blowing) downloaded from the freesound [45]. We evaluate noise starting from 50∼65 dB, with 5 dB intervals, and we play noises through an additional loudspeaker to guarantee the noise pressure level reaches the receiver at 50, 55, 60, and 65 dB. Noise samples from 4 scenes are played continuously. At the same time, we play 20 audible benign commands and deliver VRIFLE. Given that the noise is not constant, the superposition of different parts may have different effects. We repeat the above operation three times and collect 240 mixed samples for each noise level. Fig. 14 demonstrates that VRIFLE maintains effectiveness even if the noisy ambient sound reaches 65 dB with an average SR up to 97.65%. The performance drops slightly in the office noise case of 87.5%, where the keyboard typing and mouse striking are crisp noises with intense high-frequency energy. Since VRIFLE mainly affects low-frequency acoustic features after transformation, high-frequency noise might reduce its attack performance on deceiving ASR models.

6) Different Recording Devices: Since the ultrasound frequency response varies with different recording devices and microphone models [15], i.e., we establish a specific transformation model for each device. To verify that our perturbation can still manipulate the ASR model after being recorded by different devices, we obtain 5 pairs of universal and silence perturbations based on the device-wise ultrasonic transforma-

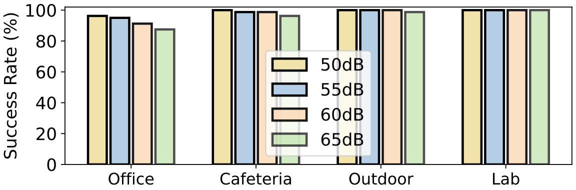


Fig. 14: Attack performance in face of noises from typical scenes at 4 sound pressure levels.

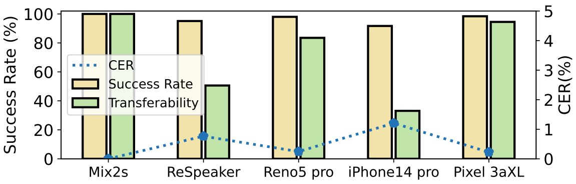


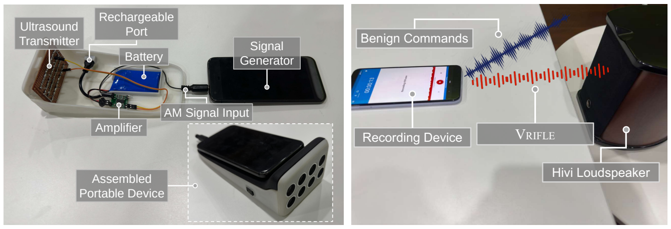
Fig. 15: Attack performance on different recording devices.

tion. After a similar collection as the above experiments done for each device, Fig. 15 depicts the average SR of these devices is up to 96.8% and CER is down to 0.50%, where Mix2s reaches 100% and 0% on these metrics, proving the crafted VRIFLE's effectiveness on individual devices. Moreover, VRIFLE gets 95.8% SR on ReSpeaker, suggesting it can also attack devices with multi-channel microphones well. Furthermore, given that adversaries may attack unmodeled (i.e., unseen) devices, we want to investigate VRIFLE's transferability despite our ultrasound transformation modeling is device-specific. We apply the optimized perturbation of Mix2s to other devices. Among them, the Mix2s' combined perturbation can transfer to Pixel 3aXL and Reno5 pro with 94.2% and 83.3% SR. Besides, the performance reduces on iPhone14 pro (31.7%) and ReSpeaker (50.8%) due to their microphones' different frequency selectivity to ultrasound. The result indicates that VRIFLE is also transferable across devices.

7) Different Speech & Perturbation Loudness: We further investigate the attack performance changes due to different loudness of the user speech and the universal perturbation. We set the representative audible sound pressure level to vary from 65∼90 dB using a decibel meter and also vary ultrasonic emission power to keep the same loudness. We play our perturbation, repeating 5 times at each volume level. Due to page limitation, results are given in Appendix §E, Fig. 20. As the mutual loudness changes, we find that once the perturbation has the same volume as the benign audio, it achieves over 55% SR. Moreover, with 5 dB higher than the benign audio, VRIFLE can work effectively with an average SR up to 95.5%. When VRIFLE's volume is 10 dB higher than audible speech, it can dominate all the user commands. Notably, even if the direct ultrasound-based attack is 35 dB louder than the audible audio, the ASR model still recognizes a CER up to 46%. In that case, VRIFLE achieves all CERs down to 0%.

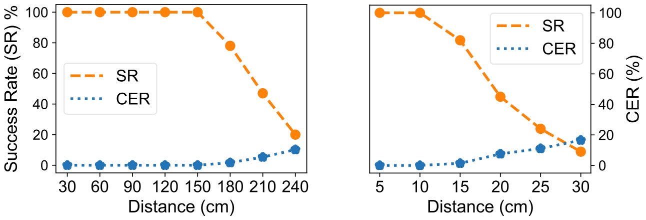
D. Attack with Portable Device and Off-the-shelf Loudspeaker

Our sophisticated device facilities an extensive attack range, providing great flexibility to attackers. We have also implemented two other covert attacks with the portable device and everyday life loudspeaker, as shown in Fig. 16.



(a) Portable Device (b) Off-the-shelf Loudspeaker

Fig. 16: Two additional attack forms of VRIFLE.



(a) Portable Device (b) Off-the-shelf Loudspeaker

Fig. 17: VRIFLE's performance at different distance with the portable device and off-the-shelf loudspeaker.

1) Portable Device: Our portable device equipped with eight 25 kHz ultrasound transducers, a compact amplifier, and a rechargeable battery in Fig. 16(a), balances lightweight and attack range. It can be connected to the smartphone, where the attacker stores perturbations as 96 kHz USB-AM audio in advance. We evaluate the effectiveness of attacks with a portable device, setting it to point at Mix2s' bottom microphone with the target "open the door". Fig. 17(a) demonstrates 100% SR within 150 cm, and 78% SR along with CER down to 1.69% even at a distance of 180 cm, suggesting VRIFLE with portable devices can exceed the attack distance of almost prior AEs.

2) Off-the-shelf Loudspeaker: Adversaries can embed USB-AM perturbations into audio or video files to manipulate user commands when played on a computer or smartphone connected to a loudspeaker. We investigate the use of off-the-shelf loudspeakers, such as the high-end Hivi [46], which have three distinct sound sources: woofer (37-140 Hz), mid-range (140-2000 Hz), and tweeter (>2000 Hz). To determine the optimal ultrasound frequency for embedding the perturbations, we conduct experiments scanning the carrier frequency from 21-27 kHz and find 25.2 kHz to be the best frequency, despite the gain decrease beyond the rated frequency range (37Hz∼20kHz). Figure17(b) illustrates that VRIFLE's effective attack distance via off-the-shelf speakers is approximately 20 cm, with a low CER of 11.07%, demonstrating effective modification of user commands at the character level.

VI. ANTI-DEFENSE EXPERIMENT

In this section, we validate whether VRIFLE can resist 6 kinds of representative defenses, involving audio preprocessing methods and inaudible attack detection. We consider two types of adversaries: 1) *Naive Adversary*: The naive adversary creates VRIFLE based on the undefended model to attack the defended model. 2) *Adaptive Adversary*: This

TABLE VI: Three defenses

(%) Method	Undefended	Quantization	VAD	Opus Compress
Success Rate	**99.49**	97.96	99.49	95.93
Attack CER	**0.1**	0.28	0.1	0.84
Benign CER	**11.23**	13.43	23.58	12.37

TABLE VII: Defense with band-pass filter

(%) Band-pass (Hz)		50∼7000	50∼6000	50∼5000	50∼4000	50∼3000
Naive Adversary	Success Rate	97.96	96.95	95.67	75.57	12.47
	Attack CER	0.28	0.53	0.80	6.43	35.01
Adaptive Adversary	Success Rate	99.75	99.24	97.20	91.60	78.88
	Attack CER	0.05	0.14	0.41	1.64	5.34
Benign CER		11.63	13.60	19.27	28.83	41.06

adversary has full knowledge of the defense mechanisms and applies customized strategies to craft VRIFLE.

Against Audio Pre-processing Defense Methods. Referring to previous works [9], [12], [31], [47] that present a series of audio pre-processing methods against audio adversarial example attacks, we examine the robustness of VRIFLE using 5 representative defenses: *(1) Quantization*: converting the audio sampling value from a 16-bit signed integer to an 8-bit precision, which reduces the sampling range from [-32,768∼32,767] to [-128∼127]. Notably, this introduces distortion and noise due to the small range of values at 8-bit precision. *(2) Voice Activity Detection (VAD)*: removing segments of audio that are less than -15 dB, where its maximum energy is normalized to 0 dB. *(3) Opus Compression Codec*: coding and compressing audio with flexible bit rate and low latency are widely used in real-time communication, particularly VoIP and online meetings. We set the default compression level as 5 according to [48]. *(4) Band-pass Filter*: filtering the input signal with given cut-off frequencies, e.g., 50∼7000 Hz. *(5) Down-sampling*: reducing the audio to a given rate, e.g., rate=0.4 means down-sampling a 16 kHz audio to 6.4 kHz, and then recovering it to the required sampling rate of targeted ASRs (generally 16/48 kHz).

We obtain the attack success rate (successfully altering the tested speech into the targeted command "open the door"), attack CER, and benign CER (derived between the DeepSpeech recognized and ground-truth transcription) at 99.49%, 0.15%, 11.23%, respectively, when the model is undefended. The results are listed in Tab. VI, VII, VIII. We observe that quantization, VAD, and Opus compression barely affect the attack success rate (all≥95.93%) in Tab. VI. Particularly, VAD significantly rises benign CER from 11.23% to 23.58%, while failing to lower our attack performance. Tab. VII and VIII demonstrate that naive VRIFLE can maintain relatively effective even when facing a 50∼4000 band-pass filter or being down-sampled to 8 kHz (rate=0.5). However, the attack performance degrades as the bandwidth or the down-sampling rate gets further lower. Note that we do not evaluate extreme cases, such as band-pass: less than 50∼2000 Hz or down-sampling rate: smaller than 0.3, since they have severely affected the model's ability to transcribe benign speech commands with unacceptable CERs over 45%. After the adaptive adversary integrates the band-pass filtering operation during optimization, the attack performance increase significantly, especially the success rate and attack CER reach 78.88% and 5.34%, respectively, even under 50∼3000 Hz band-pass filtering. Similarly, the adaptive adversary can realize an 81.42% success rate and 4.38% CER against a down-sampling rate=0.4.

TABLE VIII: Defense with down-sampling

(%) Down-sample (rate)		0.9	0.8	0.7	0.6	0.5	0.4
Naive Adversary	Success Rate	98.98	98.22	94.91	88.04	69.47	19.34
	Attack CER	0.20	0.27	1.00	2.58	8.35	30.23
Adaptive Adversary	Success Rate	99.24	98.22	96.69	95.42	89.06	81.42
	Attack CER	0.14	0.27	0.57	0.87	2.25	4.38
Benign CER		11.51	12.11	14.93	19.80	23.99	35.04

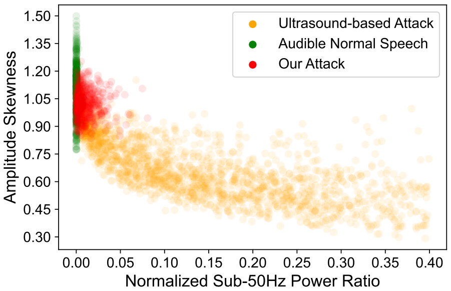


Fig. 18: Two significant feature dimensions extracted from three classes of audio samples by LipRead (800 samples/class).

Against Inaudible Attack Detection Method. Given that VRIFLE utilizes ultrasound-based modulation mechanisms, prior inaudible attack detection methods are expected to distinguish such an attack well from benign speech. We reproduce the representative software-based method: LipRead [21], strictly following its instruction, which extracts and analyzes three features of speech samples: power in sub-50Hz, correlation coefficient (between the fundamental and harmonic components), and amplitude skew. We use the LipRead classifier to detect VRIFLE samples crafted under the naive adversary setting and collected at different distances & angles; then obtain a detection accuracy down to 45.07%. Fig. 18 visualizes three types of audio samples in two significant feature dimensions. VRIFLE presents compact skewness around 1.0 due to its symmetrical waveform, whose distribution is closer to the normal, while ultrasound-based attacks appear more shift toward 0.30 and greater power in sub-50Hz. Low-frequency power aggregation is still inevitable in our attack due to nonlinear demodulation [21]. Moreover, naive VRIFLE appears low correlation coefficient compared to the traditional attacks, as its perturbations (see Fig. 7&9) barely present normal speech properties such as fundamental and harmonic frequencies. Overall, the inherent difference between VRIFLE and traditional ultrasound-based attacks makes it probably compromise LipRead. Furthermore, the adaptive adversary extracts three features during the perturbation generation and constrains them close to the normal samples, further reducing the accuracy of LipRead detecting our attack to 30.55%.

VII. Discussion and Future Work

Potential Countermeasure: We have demonstrated that VRIFLE are robust to audio pre-processing and inaudible attack detection methods. We envisage that defense approaches tracking ultrasound nature [49], [50] may be effective, although these methods are based on two hardware-dependent prototypes that can not adapt to off-the-shelf compact smart devices. For the remaining feature forensics-based [5], [21] or ML-based defenses [15], [51], we believe that the adaptive adversary shall adopt these defense strategies along with the ultrasonic transformation model during optimization and physically bypass them. But this would result in a less universal attack due to additional constraints.

Prevent Airborne Self-demodulation Leakage. Although our attack distance has significantly exceeded previous works, please note that VRIFLE cannot extend the range infinitely. As uncovered in [52], the self-demodulation occurs and then the modulated baseband becomes audible once a certain power is reached. To increase the attack range while ensuring inaudibility, we adopt the following strategies: 1) utilizing 25 kHz carrier frequency rather than higher frequencies, such as 40kHz, for less attenuation; 2) employing customized ultrasound transducers, signal generator, and amplifiers capable of suppressing nonlinear distortion at the speaker side; and 3) setting maximum power not to exceed 3.2W and implementing USB-AM to increase the attack efficiency with portable device and off-the-shelf loudspeakers.

Limitations: 1) VRIFLE achieves highly universal manipulation of user speech using DeepSpeech2's gradient information. However, its universality under black-box settings is limited in critical user-present scenarios due to variable user factors. Notably, targeted universal AE attacks in black-box scenarios remain an unsolved problem currently, despite several untargeted literature [53], [54]. 2) Although we have verified that our ultrasonic transformation model is effective on different recording devices, it is currently device-specific due to the microphone's frequency selectivity to ultrasound. We will investigate a device-generic transformation model in future work. 3) Our careful design enables the *man-in-the-middle* attack strategy and our user testing in Appendix §F demonstrate its high stealthiness. However, the testing results imply that replaying excessively long user commands may cause discomfort and might alert the user. We envision that understanding user intent and then replaying synthetic short commands can mitigate this issue.

Attack on Speaker Recognition: We envision that the idea of VRIFLE can be generalized to attack speaker recognition models deployed on access control systems, e.g., authentication of voice assistants and applications. We have conducted a preliminary experiment attacking the state-of-the-art ECAPA-TDNN [55], a popular speaker recognition model. We maintain the almost identical design as used in attacking the ASR model and only reconfigure the optimization goal y_t as the target speaker label and the loss function $\mathcal{L}(f(\cdot), y_t)$ as the cosine similarity scoring module. Results demonstrate that, in a 10-person set, VRIFLE is universal to alter the voiceprint of any user's speech samples into the targeted speaker's. We plan to delve into such an ability of VRIFLE in future work.

VIII. Related Work

Custom Adversarial Examples & Inaudible Attacks. The initial AE attacks construct a custom (i.e., non-universal) perturbation for a specific audio clip, whereas the same perturbation cannot compromise other audio. Signal-level transformations [10], [11], [56], such as modifying MFCC, are unintelligible to human beings but can be recognized by the ASR model. As this class of attacks resembles obvious noises, they can easily alert users. Thus, inaudible attacks [5], [21], [22] have been proposed, which exploit carrier signals outside the audible frequencies of human beings (e.g., 40 kHz) to inject attacks into ASR systems utilizing the nonlinearity vulnerability of microphone circuits, yet entirely unheard by victims. However, compared with audible playback speech samples, such attacks usually suffer from signal distortion and low SNR due to their dependence on various convert channels, e.g., ultrasound [57], laser [6], or electricity [24] signals, and the hardware imperfections these channels introduce. There is also a major branch of the research community that leverages the vulnerability of ASR models by adding slightly audible perturbations on the benign audio based on ϵ-constraint [7], [58] and psychoacoustic hiding [3], [4], to make the AEs sound benign but fool the ASR's transcription. It is worth noting that non-universal AEs lose effectiveness for streaming speech input and unpredictable user commands, as they rely on perfect temporal alignment. Constructing multiple AEs for altering different commands as an adversary-desired instruction is also impractical.

Universal Adversarial Examples. Recent studies propose universal AEs that can apply to tamper with multiple speech content as an adversary-desired command. Existing untargeted universal AE attacks adopt iterative greedy algorithms [59] can cause arbitrary speech to mis-classification [53] or false transcription [54]. In contrast, targeted universal AE attack is very challenging in speech recognition tasks because ASR models are context-dependent, and a certain minor perturbation superimposed even at different positions of a given benign audio, the whole sentence may yield various transcription results. This is distinct from the prior successful targeted universal AE attack in the text-independent speaker recognition [26], [31] and the universal adversarial patch attack in position-insensitive CNN-based image classification tasks [60]. Moreover, given that the victim user can easily notice the audible-band perturbation, AdvPulse [9] disguises short pulses in the environment sounds to be less perceptible. However, they only apply to a context-insensitive CNN-based audio command classification model to be universal. To overcome the mainstream RNN-based ASR context-dependent issues, a partial match strategy is proposed by SpecPatch [8], which also employs audible noise-like short pulses (0.5s) to alter multiple short user commands into the targeted instruction against the mainstream DeepSpeech ASR model. However, such an attack will not work in relatively long commands ($\geq$ 4 words) and can be noticed despite following L2-imperceptibility constraints.

Overall, due to the fundamental differences between audible and ultrasonic channels, VRIFLE differs from prior works that encountered challenges related to ***user auditory*** and ***user disruption***. In addition to the four representative merits over existing AEs listed in Tab. I, VRIFLE offers several additional benefits: (1) the optimization process is no longer subject to audibility constraints such as tiny ϵ, psychoacoustics, L_p-norm, nor does it need to limit the signal form as short pulses to reduce the possibility of being perceived. (2) VRIFLE's broad optimization space further allows for fewer iterations while maintaining a high degree of universality. Combining these two advantages, VRIFLE enables real-time manipulation of arbitrary user commands and long speech sentences in an *alter-and-mute* fashion, as never before. (3) Unlike audible-band AEs that are easily compromised by interference due to their subtle perturbations, VRIFLE demonstrates robustness and remains effective even when faced with various audio pre-processing defenses. Notably, our initial modeling of ultrasound transformation precisely characterizes the ultrasound channel and justifies it as a promising carrier for IAP delivery. We believe that this modeling effort lays the groundwork for generating inaudible AEs and may inspire future works.

IX. Conclusion

In this work, we propose an inaudible adversarial perturbation (IAP) attack against ASR systems named VRIFLE, which can extend to scenarios where users are present and may use ASR services. In such scenarios, prior studies will fail due to *user auditory* and *user disruption*. We make the first attempt to model the ultrasonic transformation process, based on which, VRIFLE can alter arbitrary user commands to the adversary-desired intent in real time without any knowledge of users' speech. Our comprehensive experiments in the digital and physical worlds across various configurations demonstrate VRIFLE's effectiveness and robustness. Overall, VRIFLE features merits including complete inaudibility, universality, and long-range attack ability.

Acknowledgement

We thank the shepherd and anonymous reviewers for their valuable comments and suggestions. We also appreciate Dr. Jiangyi Deng for providing valuable insights and guidance to the paper's content. This work is supported by China NSFC Grant 61925109, 62201503, 62222114, and 62071428.

References

[1] Microsoft Azure. Azure speech-to-text. https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/, 2021.

[2] Google Speech. Google cloud speech-to-text. https://cloud.google.com/speech-to-text/, 2021.

[3] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. *arXiv preprint arXiv:1808.05665*, 2018.

[4] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5231–5240, 2019.

[5] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 103–117, 2017.

[6] Takeshi Sugawara, Benjamin Cyr, Sara Rampazzi, Daniel Genkin, and Kevin Fu. Light commands: laser-based audio injection attacks on voice-controllable systems. In *Proceedings of the 29th USENIX Security Symposium (USENIX Security 20)*, pages 2631–2648, 2020.

[7] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)*, pages 1–7. IEEE, 2018.

[8] Hanqing Guo, Yuanda Wang, Nikolay Ivanov, Li Xiao, and Qiben 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*, pages 1353–1366, 2022.

[9] Zhuohang Li, Yi Wu, Jian Liu, Yingying Chen, and Bo 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*, pages 1121–1134, 2020.

[10] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In *25th USENIX security symposium (USENIX security 16)*, pages 513–530, 2016.

[11] Hadi Abdullah, Washington Garcia, Christian Peeters, Patrick Traynor, Kevin R. B. Butler, and Joseph Wilson. Practical hidden voice attacks against speech and speaker recognition systems. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019*, 2019.

[12] Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, Xiaofeng Wang, and Carl A. Gunter. Commandersong: A systematic approach for practical adversarial voice recognition. In *27th USENIX Security Symposium, USENIX Security*, pages 49–64, 2018.

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

[14] Guangke Chen, Sen Chen, Lingling Fan, Xiaoning Du, Zhe Zhao, Fu Song, and Yang Liu. Who is real bob? adversarial attacks on speaker recognition systems. In *42nd IEEE Symposium on Security and Privacy, SP 2021*, pages 694–711. IEEE, 2021.

[15] Xinfeng Li, Xiaoyu Ji, Chen Yan, Chaohao Li, Yichen Li, Zhenning Zhang, and Wenyuan Xu. Learning normality is enough: A software-based mitigation against the inaudible voice attacks. In *Proceedings of the 32nd USENIX Security Symposium*, 2023.

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

[17] Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

[18] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:, 1994.

[19] Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012.

[20] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016.

[21] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. Inaudible voice commands: The long-range attack and defense. In *Proceedings of the 15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*, pages 547–560, 2018.

[22] Qiben Yan, Kehai Liu, Qin Zhou, Hanqing Guo, and Ning Zhang. Surfingattack: Interactive hidden attack on voice assistants using ultrasonic guided waves. In *Proceedings of the Network and Distributed Systems Security (NDSS) Symposium*, 2020.

[23] Xiaoyu Ji, Juchuan Zhang, Shui Jiang, Jishen Li, and Wenyuan Xu. Capspeaker: Injecting voices to microphones via capacitors. In *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea*, pages 1915–1929, 2021.

[24] Yuanda Wang, Hanqing Guo, and Qiben Yan. Ghosttalk: Interactive attack on smartphone voice system through power line. In *Network and Distributed Systems Security (NDSS) Symposium*, 2022.

[25] Nirupam Roy, Haitham Hassanieh, and Roy Choudhury. Backdoor: Making microphones hear inaudible sounds. In *Proceedings of the 15th*

Annual International Conference on Mobile Systems, Applications, and Services, pages 2–14, 2017.

[26] Xinfeng Li, Junning Ze, Chen Yan, Yushi Cheng, Xiaoyu Ji, and Wenyuan Xu. Enrollment-stage backdoor attacks on speaker recognition systems via adversarial ultrasound. *arXiv preprint arXiv:2306.16022*, 2023.

[27] Avisoft Bioacoustics. Ultrasonic dynamic speaker vifa. http://www.avisoft.com/usg/vifa.htm, 2017.

[28] Tmall. Tct25-16t 25 khz ultrasonic transducer. https://item.taobao.com/item.htm?spm=a230r.1.14.1.3e5969e6Wy38Kn&id=539957350166&ns=1&abbucket=12#detail, 2023.

[29] fluent.ai. Fluent speech commands. Website, 2020. https://www.kaggle.com/tommyngx/fluent-speech-corpus.

[30] Steve TK Jan, Joseph Messou, Yen-Chen Lin, Jia-Bin Huang, and Gang Wang. Connecting the digital and physical world: Improving the robustness of adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 962–969, 2019.

[31] Jiangyi Deng, Yanjiao Chen, and Wenyuan Xu. Fencesitter: Black-box, content-agnostic, and synchronization-free enrollment-phase attacks on speaker recognition systems. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 755–767, 2022.

[32] Hiromu Yakura and Jun Sakuma. Robust audio adversarial example for a physical attack. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, 2019*, pages 5334–5341, 2019.

[33] Marco Jeub, Magnus Schafer, and Peter Vary. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *2009 16th International Conference on Digital Signal Processing*, pages 1–5. IEEE, 2009.

[34] Satoshi Nakamura, Kazuo Hiyane, Futoshi Asano, Takanobu Nishiura, and Takeshi Yamada. Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, 2000.

[35] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement. *arXiv preprint arXiv:2008.00264*, 2020.

[36] Angelo Farina. Advancements in impulse response measurements by sine sweeps. In *Audio engineering society convention 122*. Audio Engineering Society, 2007.

[37] Sverre Holm. *Waves with power-law attenuation*. Springer, 2019.

[38] Rodger Ziemer and William H Tranter. *Principles of communications: system modulation and noise*. John Wiley & Sons, 2006.

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[40] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1, 2014.

[41] Siglent. Sdg6032x-e. Website, 2019. https://siglentna.com/product/sdg6032x/.

[42] Micronix. Nf hsa4015. Website, 2013. https://eshop.micronix.eu/measurement-equipment/electrical-quantities/nf-corporation-instruments/high-speed-bipolar-amplifiers/hsa-4051.html.

[43] Seeed Studio. Respeaker mic array v2.0. https://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/, 2021.

[44] Giuseppe Petracca, Yuqiong Sun, Trent Jaeger, and Ahmad Atamli. Audroid: Preventing attacks on audio channels in mobile devices. In *Proceedings of the 31st Annual Computer Security Applications Conference*, pages 181–190, 2015.

[45] freesound. freesound.org. https://freesound.org/, 2022.

[46] Hivi. Swan m-50wmkiii. Website. https://www.swanspeakers.com/product/view?id=162.

[47] Shehzeen Hussain, Paarth Neekhara, Shlomo Dubnov, Julian McAuley, and Farinaz Koushanfar. Waveguard: Understanding and mitigating audio adversarial examples. *arXiv preprint arXiv:2103.03344*, 2021.

[48] Facebook. Torchaudio functional. https://pytorch.org/audio/master/functional.html#torchaudio.functional.apply_codec, 2022.

[49] Guoming Zhang, Xiaoyu Ji, Xinfeng Li, Gang Qu, and Wenyuan Xu. Eararray: Defending against dolphinattack via acoustic attenuation. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, 2021.

[50] Yitao He, Junyu Bian, Xinyu Tong, Zihui Qian, Wei Zhu, Xiaohua Tian, and Xinbing Wang. Canceling inaudible voice commands against voice control systems. In *Proceedings of the 25th Annual International Conference on Mobile Computing and Networking*, pages 1–15, 2019.

[51] Zhuohang Li, Cong Shi, Tianfang Zhang, Yi Xie, Jian Liu, Bo Yuan, and Yingying Chen. Robust detection of machine-induced audio attacks in intelligent audio systems with microphone array. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 1884–1899, 2021.

[52] Ryo Iijima, Shota Minami, Zhou Yunao, Tatsuya Takehisa, Takeshi Takahashi, Yasuhiro Oikawa, and Tatsuya Mori. Audio hotspot attack: An attack on voice assistance systems using directional sound beams. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 2222–2224, 2018.

[53] Jon Vadillo and Roberto Santana. Universal adversarial examples in speech command classification. *arXiv preprint arXiv:1911.10182*, 2019.

[54] Paarth Neekhara, Shehzeen Hussain, Prakhar Pandey, Shlomo Dubnov, Julian McAuley, and Farinaz Koushanfar. Universal adversarial perturbations for speech recognition systems. *arXiv preprint arXiv:1905.03828*, 2019.

[55] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association*, pages 3830–3834, 2020.

[56] Tavish Vaidya, Yuankai Zhang, Micah Sherr, and Clay Shields. Cocaine noodles: exploiting the gap between human and machine speech recognition. In *9th USENIX Workshop on Offensive Technologies (WOOT 15)*, 2015.

[57] Xiaoyu Ji, Juchuan Zhang, Shui Jiang, Jishen Li, and Wenyuan Xu. Capspeaker: Injecting voices to microphones via capacitors. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021.

[58] Rohan Taori, Amog Kamsetty, Brenton Chu, and Nikita Vemuri. Targeted adversarial examples for black box audio systems. In *2019 IEEE security and privacy workshops (SPW)*, pages 15–20. IEEE, 2019.

[59] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[60] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

APPENDIX

A. SINGLE-SIDEBAND AMPLITUDE MODULATION

In this section, we give mathematical proof that the baseband perturbation of SSB-AM signals can be recovered by commercial microphones. We initially compare the maximum energy of USB-AM and LSB-AM emitting the same perturbation when sound leakage occurs, and LSB-AM is 87% of USB-AM. Thus, we adopt the USB-AM in our attacks due to its better inaudibility:

USB-AM: $S_{USB}(t) = mcos\omega_c t - \hat{m}sin\omega_c t + cos\omega_c t$

LSB-AM: $S_{LSB}(t) = mcos\omega_c t + \hat{m}sin\omega_c t + cos\omega_c t$

where the $\hat{m}$ is the conjugate of m. The microphone amplifier's output is below:

$$S_{out} = k_1 S_{USB}(t) + k_2 S_{USB}^2(t) + \cdots$$

The $S_{USB}^2(t)$ term has three components: a high-frequency $2\omega_c t$ components:

$$(m+1)\hat{m}\sin(2\omega_c t) + \frac{m^2+2m+1-\hat{m}^2}{2}\cos(2\omega_c t)$$

a direct current (DC) term $\frac{1}{2}$ and an audible component $S_{aud}(t) = \frac{1}{2}(m^2+2m+\hat{m}^2)$. $S_{USB}(t)$ and the high-frequency component are filtered by the low-pass filter because its frequency is above 25 kHz. The DC component is filtered by the microphone's capacitor. Thus, the audible component $S_{aud}(t)$ that passes the microphone filtering system can function to ASR.

B. Real-world Scenario

Figure 19 presents our real-world attack scenario.

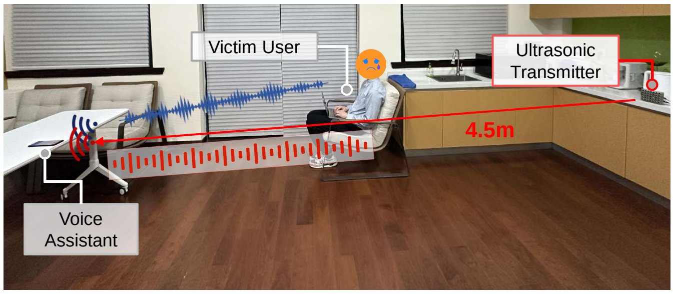


Fig. 19: Real-World Attack Scenario.

C. Targeted Commands Lists

Tab.IX lists 10 different commands, corresponding to the performance of constructing target command-specific perturbations in experiment §V-B3.

TABLE IX: Attack with Different Targeted Commands

Target Command	SR	CER
"Start recording"	100%	0%
"Set a timer"	100%	0%
"Open the door"	100%	0%
"Take the picture"	100%	0%
"Call nine one one (911)"	100%	0%
"Cancel my morning alarm"	100%	0%
"Turn on airplane mode"	94.39%	0.28%
"Open my photo album"	95.03%	0.50%
"What is going on Twitter?"	100%	0%
"Mute volume and turn off the WiFi"	92.82%	0.21%

D. Algorithm of Vrifle

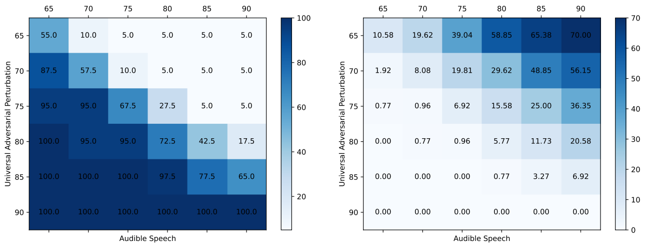
Given that technical workflow for the silence and universal perturbation are overall identical, the major differences are the optimization objective: y_t/y_b and hyper-parameters. Therefore, we demonstrate Vrifle's representative optimization process of crafting a universal perturbation from scratch in Algorithm. 1.

Algorithm 1: Universal Vrifle Generation

```
Input: The ASR model with CTC Loss Computation
       module: $\mathcal{L}$, the maximum epoch: maxEpoch,
       the desired loss: $objValue$, with a scoring
       module: $S$, the learning rate: $\eta$, the preset time
       range: $T$.
Output: The universal perturbation $\delta$
Init $\delta \leftarrow 0^N$
for 1 to $maxEpoch$ do
    $J \leftarrow 0$
    for $h_\theta \in U_H,\ n \in U_N$ do
        $\hat{e} = e^{-a_0 \omega_c^n d}$
        $\overline{\delta} = h_\theta \hat{e} * \overline{\delta : \hat{\xi}} + n$
        for $x \in U_x, g \in G, S_{(\cdot)}$ s.t. $T$ do
            $\tilde{x} = \beta \cdot g * x$
            $\tilde{x_\delta} = clip(\tilde{x} + \mathcal{S}_{(\overline{\delta})}, [-1, 1])$
            $J += \mathcal{L}(\tilde{x_\delta}, y_t)$
        end
    end
    Compute $\nabla_\delta J$
    $\delta \leftarrow \Omega_{Adam}(\delta + \eta \cdot \nabla_\delta J)$
    $\delta \leftarrow clip(\delta, [-1, 1])$
    if $J \leq objValue$ then
        break
    end
end
```

E. Different Speech & Perturbation Loudness

Fig. 20 shows the success rate and CER of our experiments on the relative energies between the attack perturbation and speech.



(a) Success Rate (%)

(b) CER (%)

Fig. 20: The performance of loudness relationship between user speech and perturbation.

F. User Testing

In this section, we elaborate on the *Man-in-the-middle* attack strategy, whose effect is akin to experiencing network congestion when users use the ASR service, resulting in slower responses. Prolonged latency can make users feel uncomfortable while using the service. To assess user awareness under such delays, we design 10 scenarios, each consisting of an audio clip that simulates a user issuing a command to the ASR system with random delays (1-5 seconds) before the voice assistant executes the command. We collected test results from 140 college students of different majors. As shown in Figure 21, when the delay time is less than 2.7 seconds (the junction point of two distribution curves), more users find the ASR service comforting than uncomfortable. The participants are also asked to fill in what they think the cause is if they experience an uncomfortable delay when using the ASR service. Only 11 out of 140 participants suspect an attack, while almost all others attribute the delay to network latency/congestion or device stuck, suggesting that this strategy poses a hidden attack. We believe that users' suspicion may also be related to their disciplinary background, e.g., users with knowledge of cybersecurity are more likely to consider the possibility of an attack.

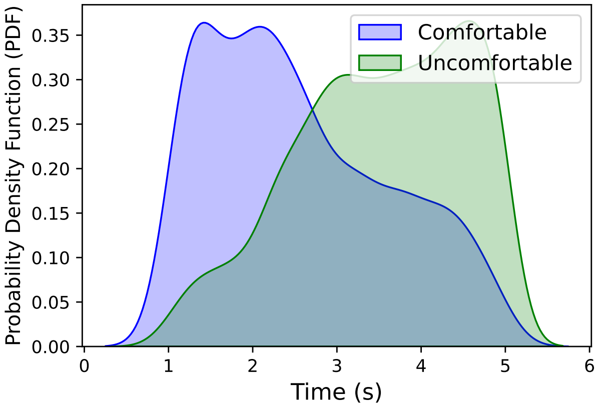


Fig. 21: The probability distribution of users' awareness during a *man-in-the-middle* attack under different delay conditions (similar to network latency). "Comfortable": the situation where users find the ASR service is normal and are not aware of the attack; "Uncomfortable": the delay may cause them to feel uncomfortable or unusual.