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

Xinfeng Li, Chen Yan†, Xuancun Lu, Zihan Zeng, Xiaoyu Ji†, Wenyuan Xu

Ubiquitous System Security Lab (**USSLAB**), Zhejiang University

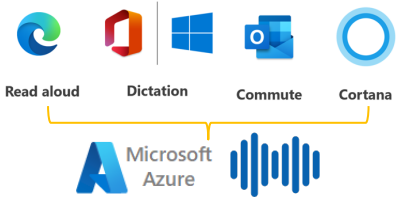





Automatic Speech Recognition (ASR) are Everywhere!

Apple Siri

Microsoft Azure



Google Home

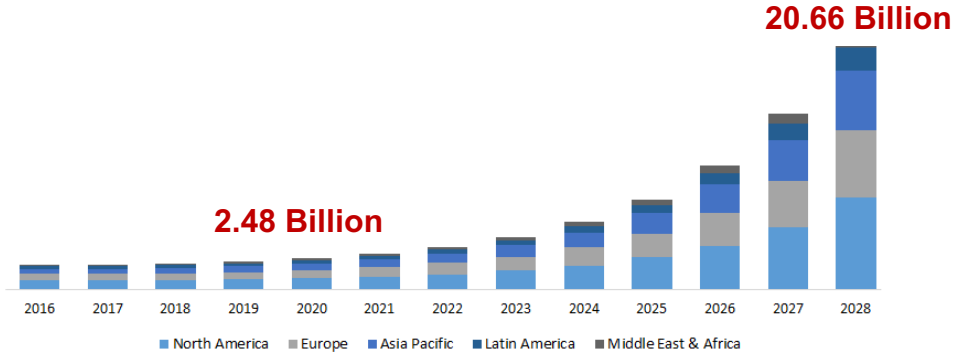
Amazon Echo

Read my message

Call my boss

Open the door

Voice Assistant Application Market Size, By Region, 2016 - 2028
(USD Billion)

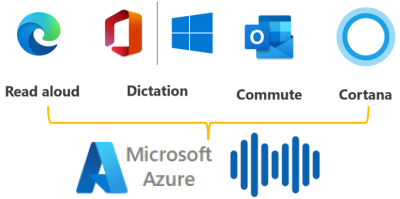


Source: Polaris Market Research Analysis

Automatic Speech Recognition (ASR) are Everywhere!

Apple Siri

Microsoft Azure



Google Home

Amazon Echo

Read my message

Call my boss

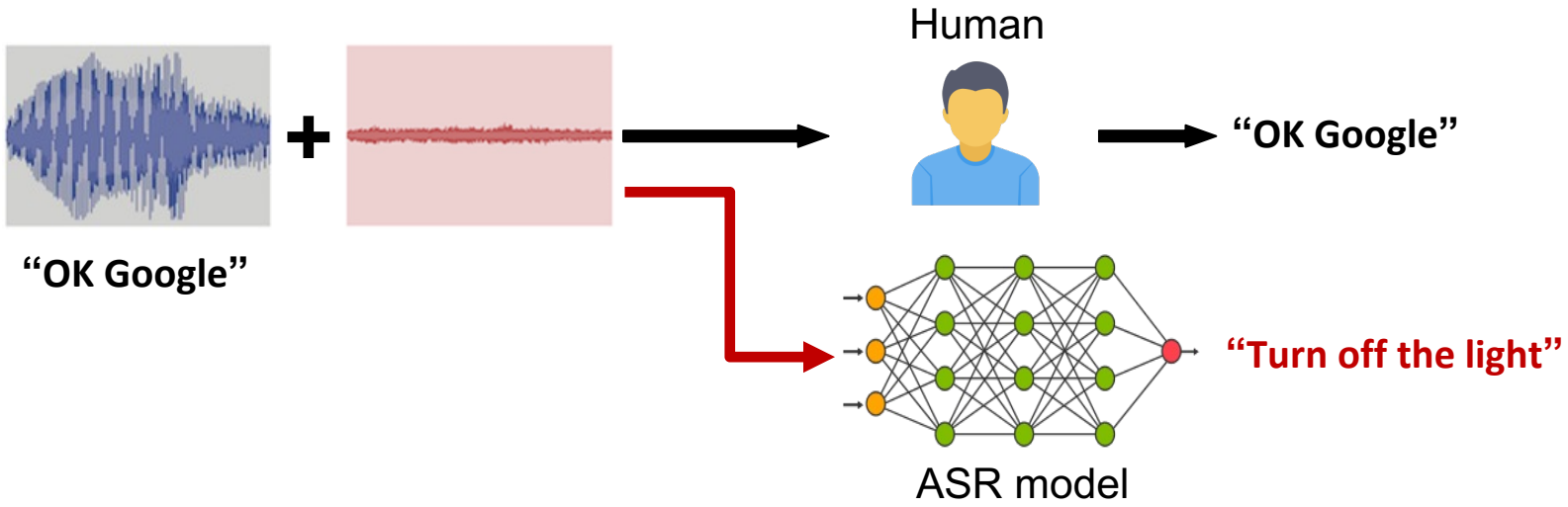
Open the door

Adversarial Example

Source: Polaris Market Research Analysis

Audio Adversarial Examples against ASRs

- AE does not impact human comprehension, while spoofing ASR models

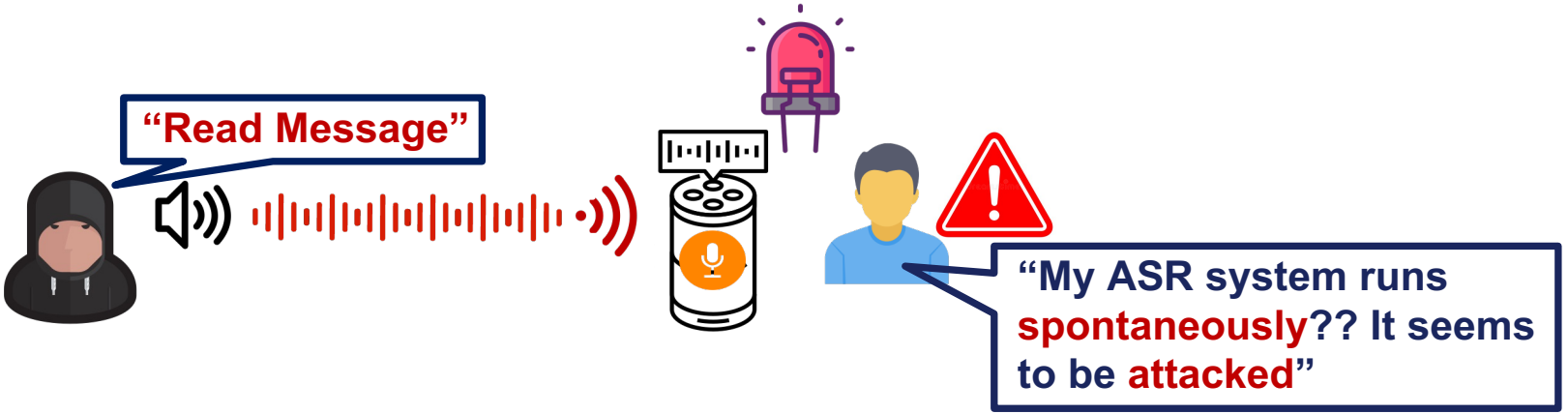


Threat Model: User-present Scenario

- ASR systems always respond with Vocal Prompt / LED blink once receiving commands

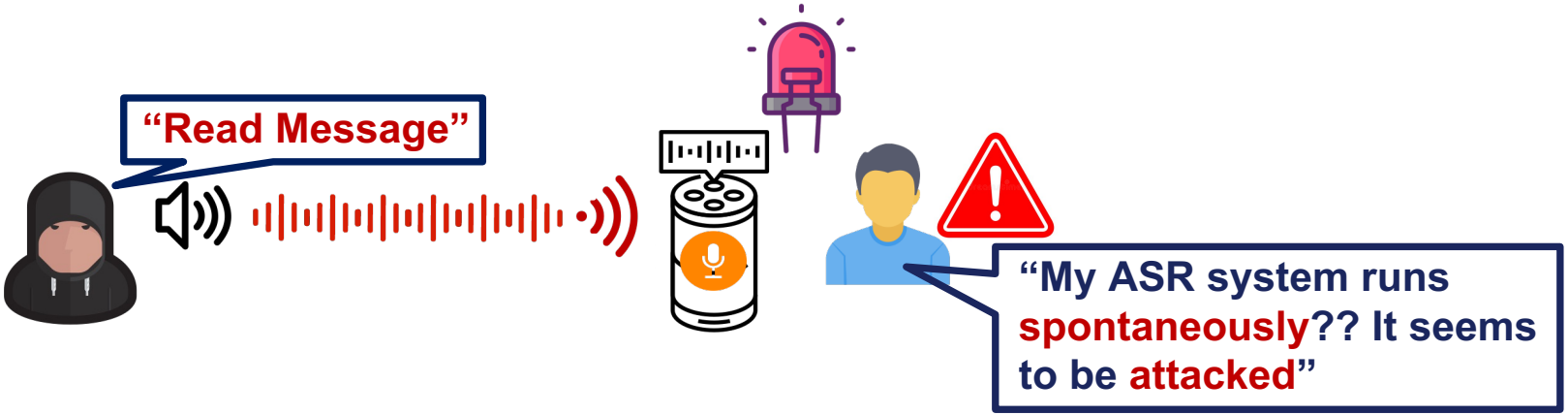
Threat Model: User-present Scenario

- ASR systems always respond with Vocal Prompt / LED blink once receiving commands



Threat Model: User-present Scenario

- ASR systems always respond with Vocal Prompt / LED blink once receiving commands



How to attack ASR systems while avoiding alerting users?

Threat Model: User-present Scenario

- **Attack when users are speaking**, as they are expect the ASR system's reaction, attack results are less suspicious

Ideal AE Attack Properties

- **Stealthy**

Ideal AE Attack Properties

- **Stealthy**

- **Universal**

Ideal AE Attack Properties

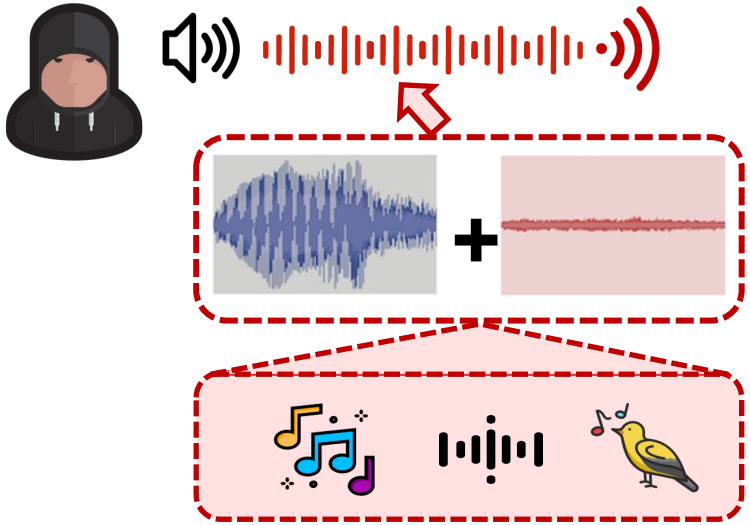
- **Stealthy**

- **Universal**

- **Practical**

Prior Attack Limitations

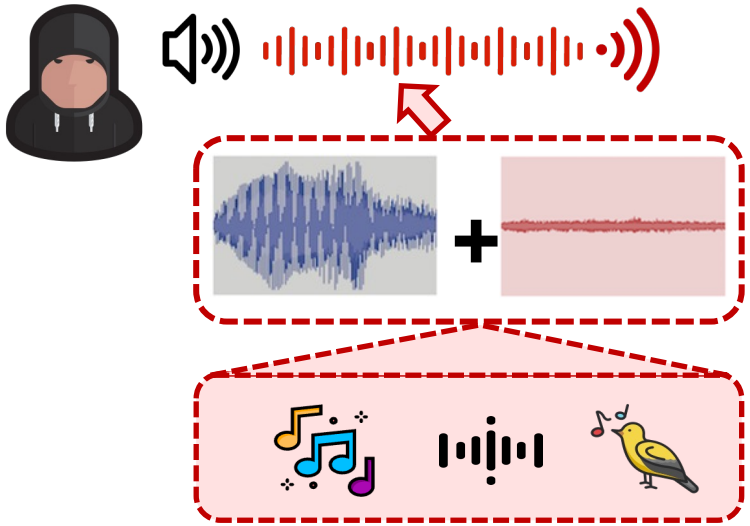
- Despite employing stealthiness constraints to limit perturbations small or hide them into innocent sounds.



- Music-like
- Noise
- Ambient Sound
- Short Pulse

Prior Attack Limitations

- Despite employing stealthiness constraints to limit perturbations small or hide them into innocent sounds.



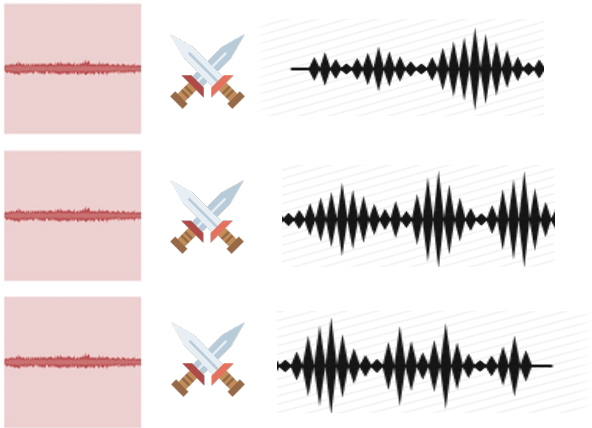
"I noticed suspicious and continue music/noise…"

- Music-like
- Noise
- Ambient Sound
- Short Pulse

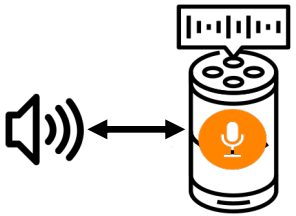
Prior Attack Limitations

- Stealthiness constraint limits prior AEs's universality and practicality.

Universal



Practical

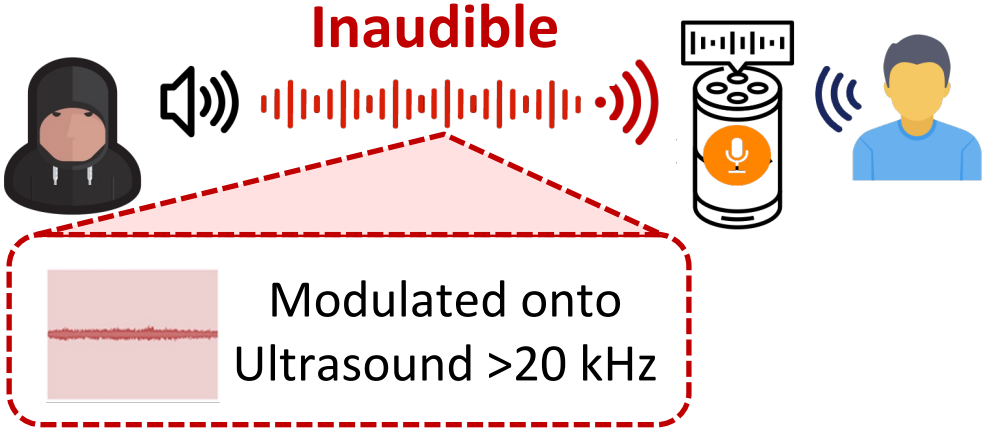


Normally less than 2 meter

Can we achieve inaudible & universal & practical AE attack to manipulate user speech in real time?

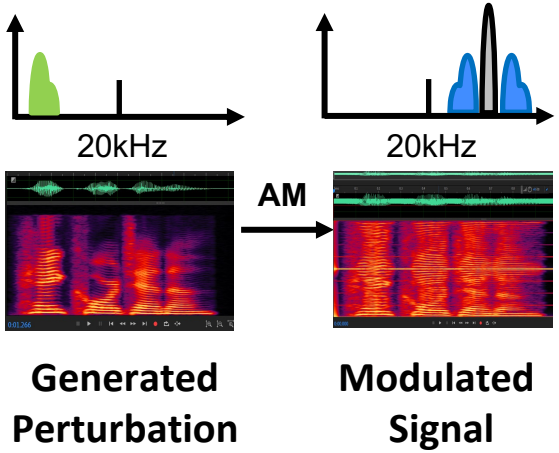
Vrifle: Basic Idea

- Completely inaudible to human beings via ultrasonic delivery



Vrifle: Rationale of Ultrasound-based Delivery

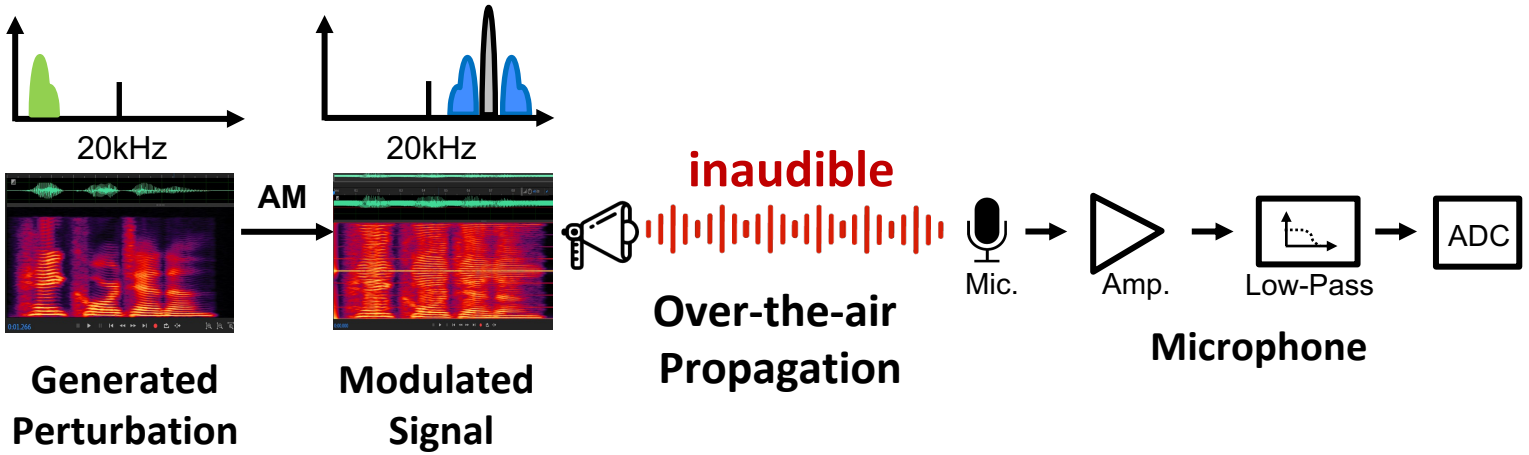
- How can Vrifle achieve inaudible delivery?



[1] Zhang, Guoming, 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*, pp. 103-117. 2017.

Vrifle: Rationale of Ultrasound-based Delivery

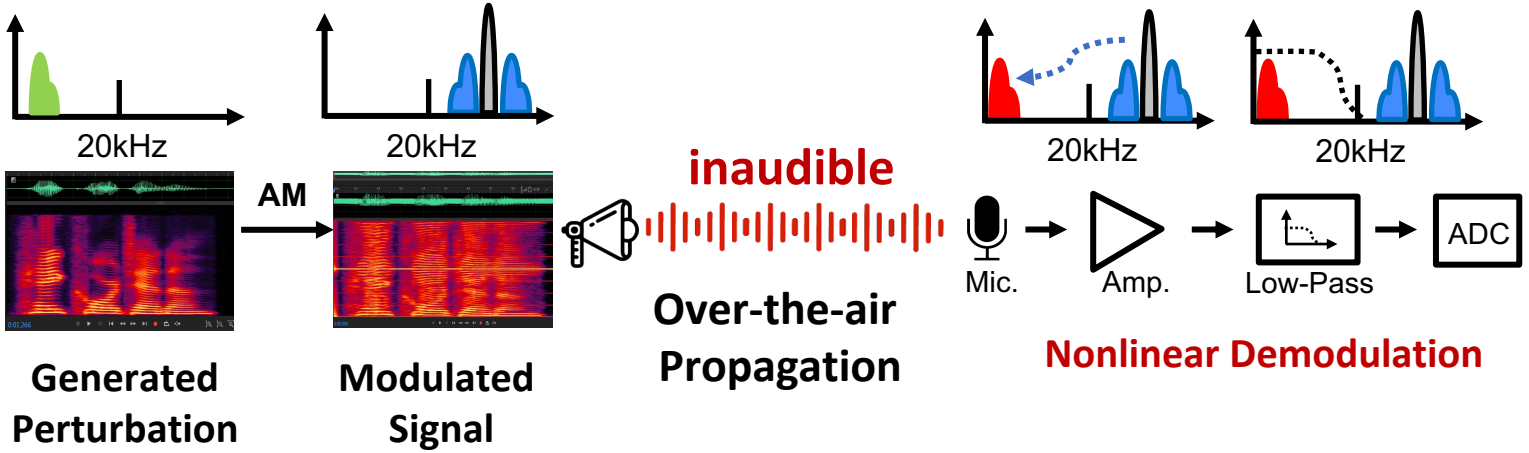
➢ How can Vrifle achieve inaudible delivery?



[1] Zhang, Guoming, 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*, pp. 103-117. 2017.

Vrifle: Rationale of Ultrasound-based Delivery

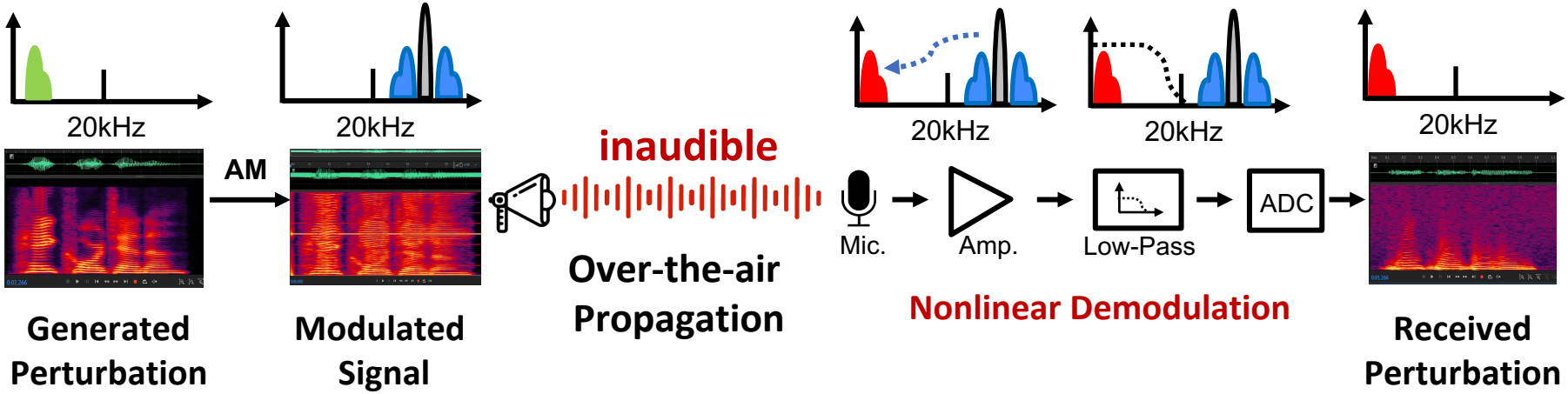
➢ How can Vrifle achieve inaudible delivery?



[1] Zhang, Guoming, 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*, pp. 103-117. 2017.

Vrifle: Rationale of Ultrasound-based Delivery

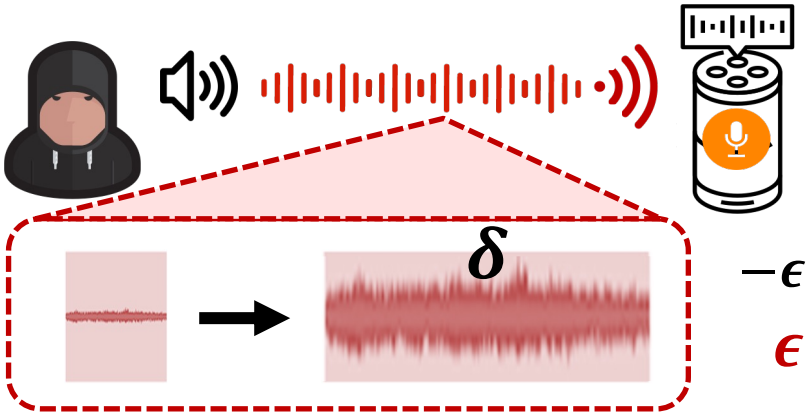
➢ How can Vrifle achieve inaudible delivery?



[1] Zhang, Guoming, 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*, pp. 103-117. 2017.

Vrifle: Real-Time Manipulation of User Speech

- Remove stealthiness constraints conflict with universality

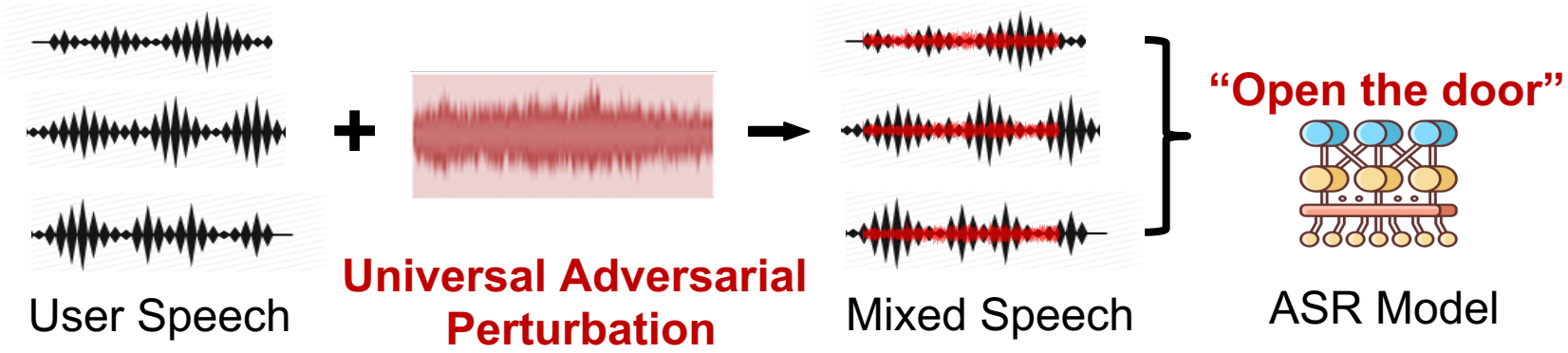


$$-\epsilon \leq |\delta| \leq \epsilon$$

$\epsilon = 1.0$ (1.0 is the upper bound)

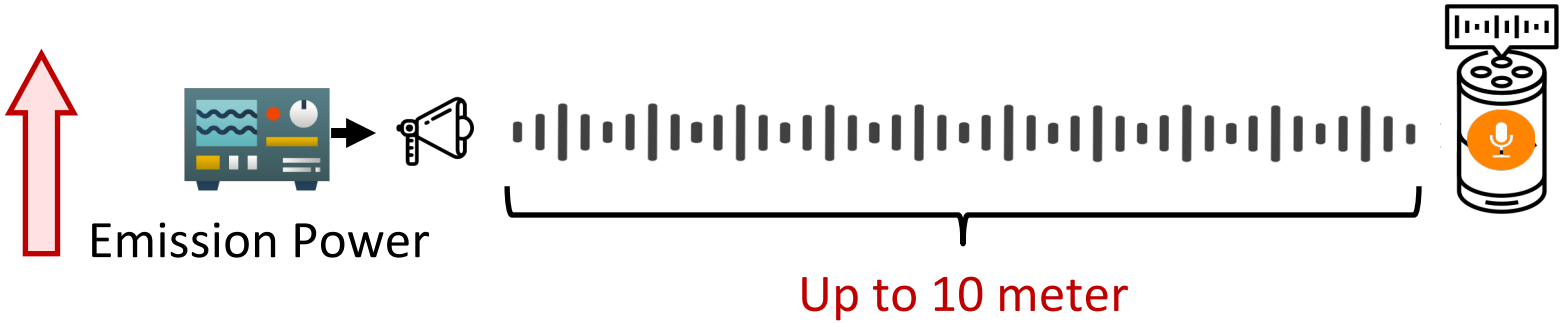
Vrifle: Real-Time Manipulation of User Speech

- Universal to tamper with any user speech

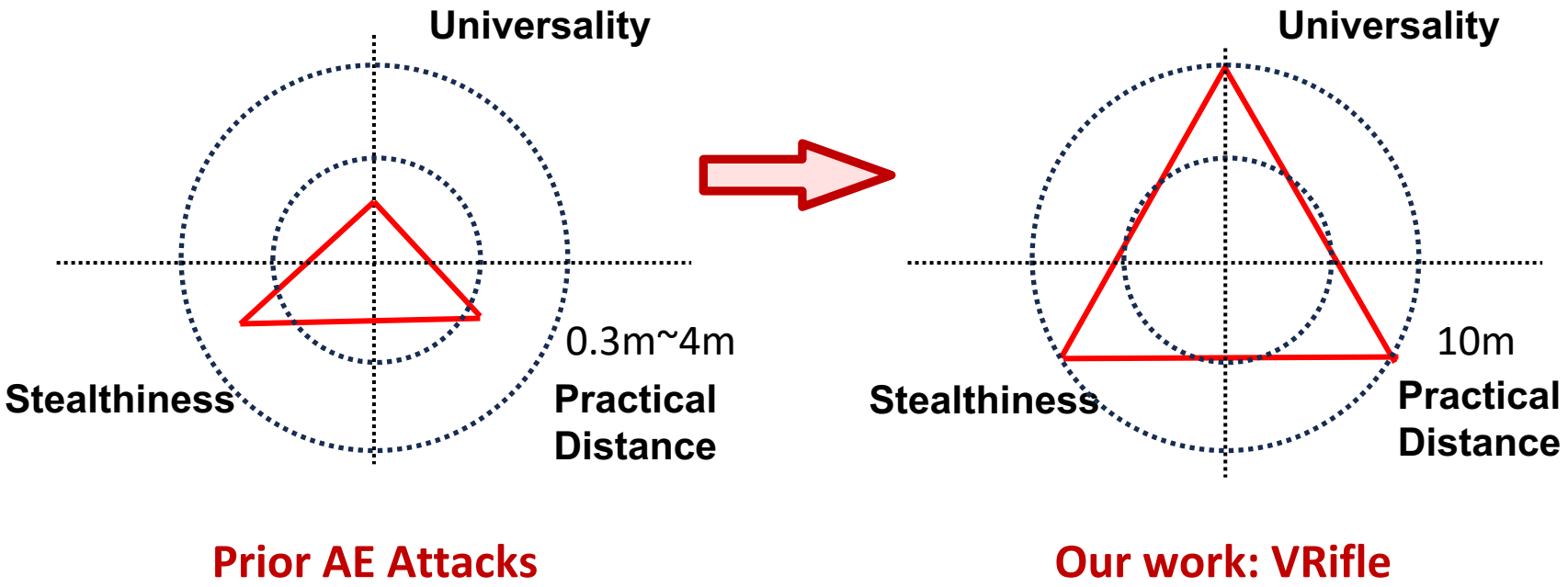


Vrifle: Real-Time Manipulation of User Speech

- Practical to achieve long-range attack



New Paradigm: Inaudible & Universal & Practical



Prior AE Attacks

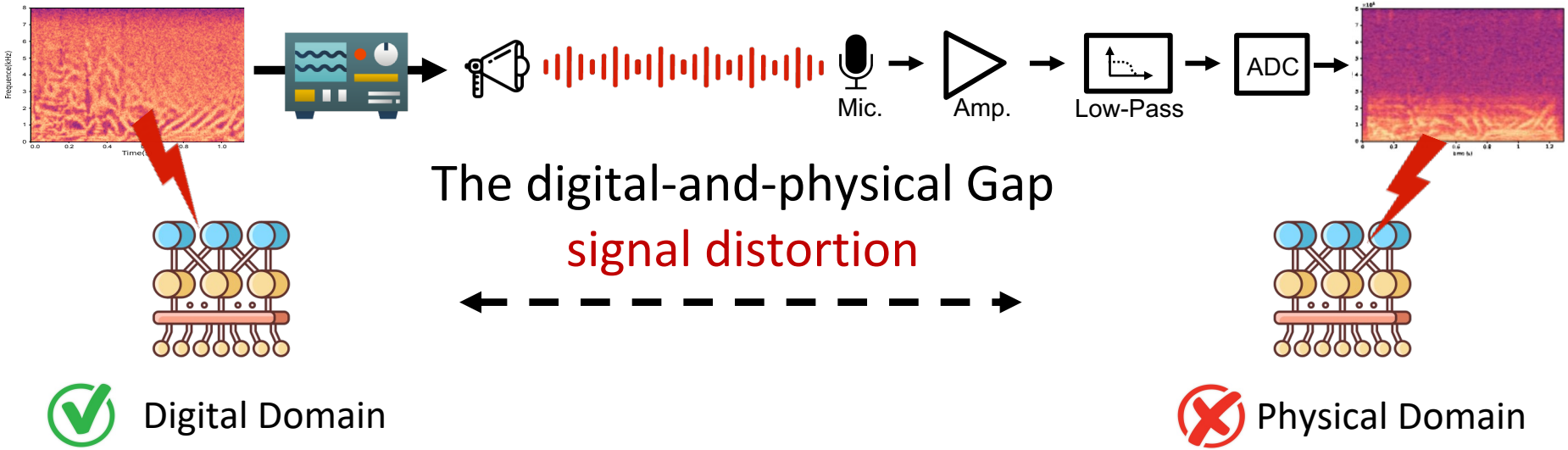
Our work: VRifle

Key Challenges

- ***Practical Challenge---Perturbation Distortion:*** Fine-grained AEs are physically ineffective after long-range propagation and complex transformation.
- ***Universal Challenge---Unpredictable User Speech****:* A fixed-length perturbation cannot tamper with excessively long user speech
- ***Equipment Challenge---Sound Leakage:*** Ultrasound delivered by unspecialized device may lead to sound leakage.

Challenge1: Perturbation Distortion

- AE works in digital domain but is ineffective in physical world



Challenge1: Perturbation Distortion

- RIR and ML-based methods are not applicable

Room Impulse Response ① (RIR)

Apply for audible-band AEs

Cannot estimate AEs delivered via Ultrasound

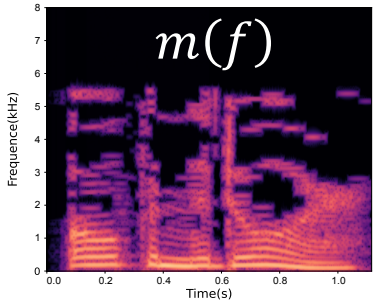
End-to-End ② ML-based Modeling

Require lots of paired data

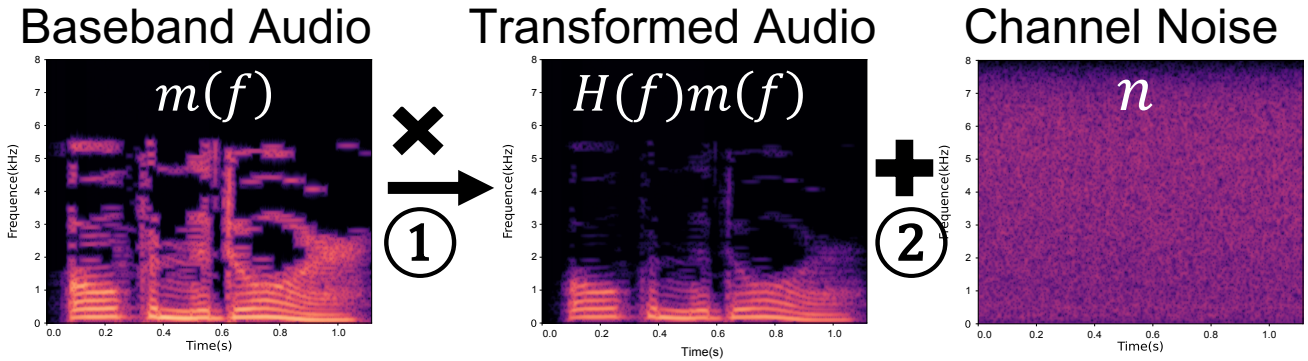
Hard to enable location-free attacks

Solution: Ultrasonic Transformation Modeling

Baseband Audio

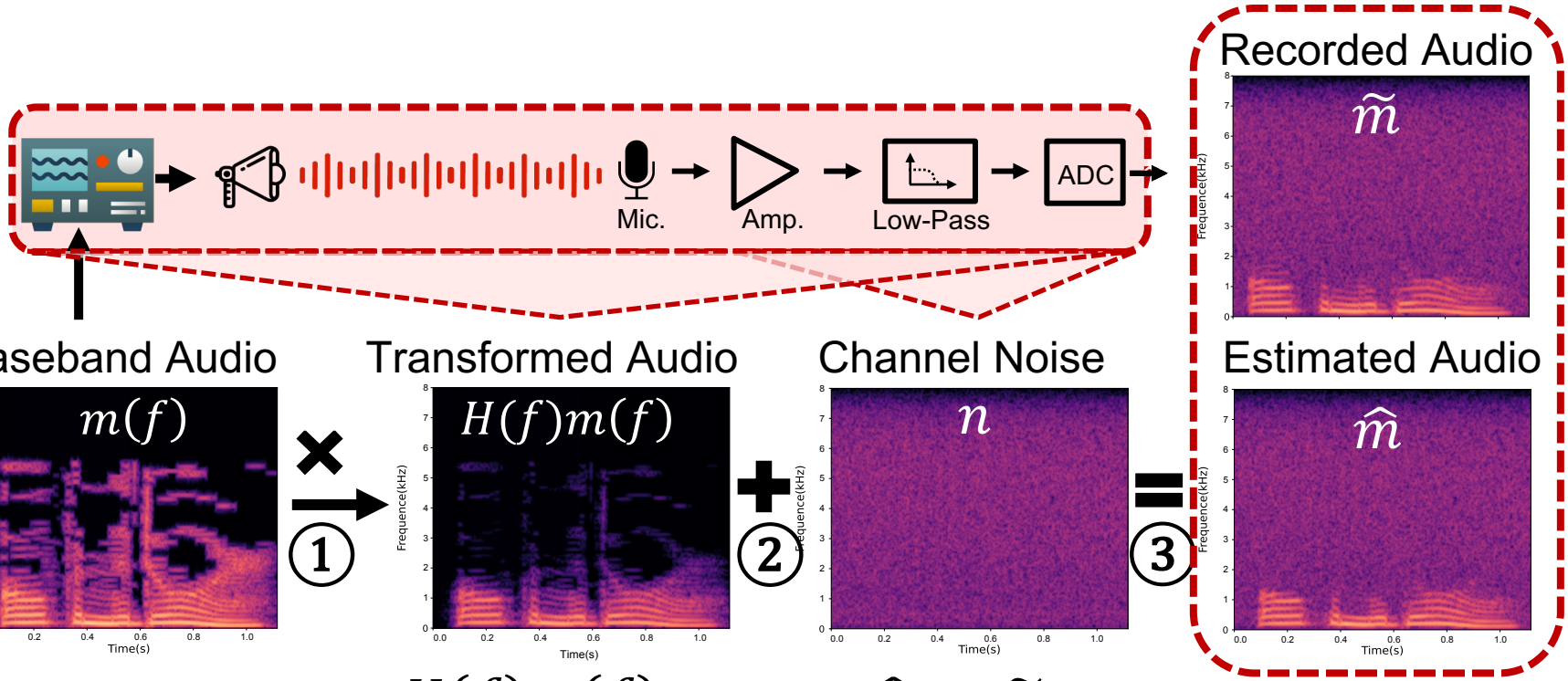


Solution: Ultrasonic Transformation Modeling



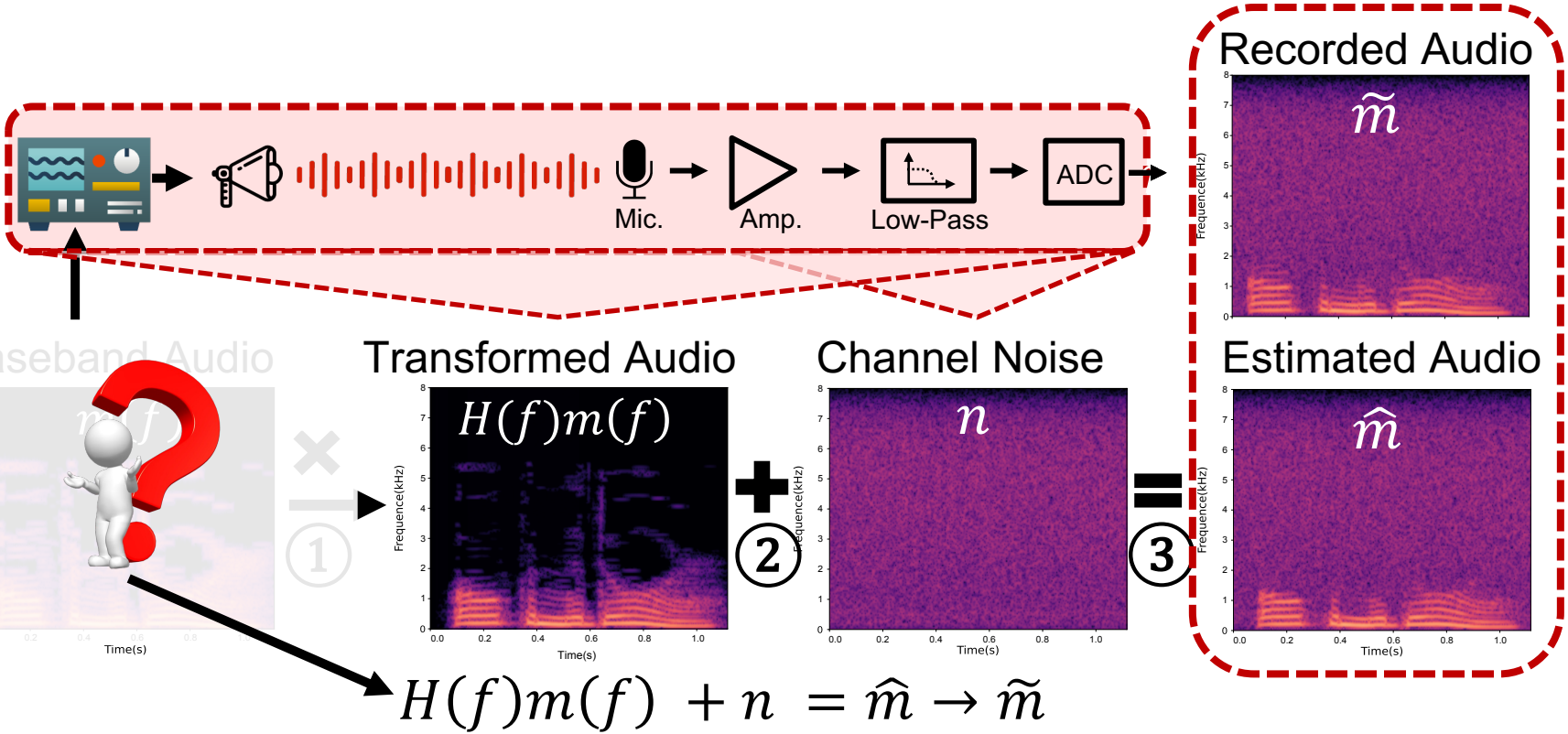
$$H(f)m(f) + n = \hat{m} \rightarrow \tilde{m}$$

Solution: Ultrasonic Transformation Modeling



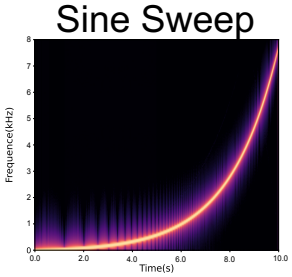
$$H(f)m(f) + n = \hat{m} \rightarrow \tilde{m}$$

Solution: Ultrasonic Transformation Modeling



Solution: Ultrasonic Transformation Modeling

- Step-by-step derive ultrasound frequency response (UFR)

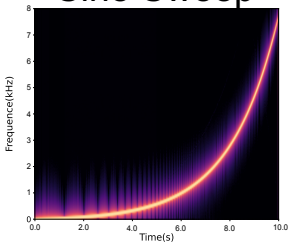


Generate 10-second Sine Sweep

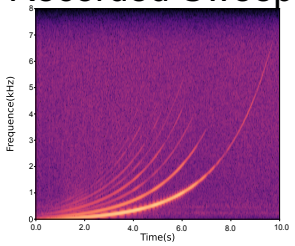
Solution: Ultrasonic Transformation Modeling

- Step-by-step derive ultrasound frequency response (UFR)

Sine Sweep



Recorded Sweep

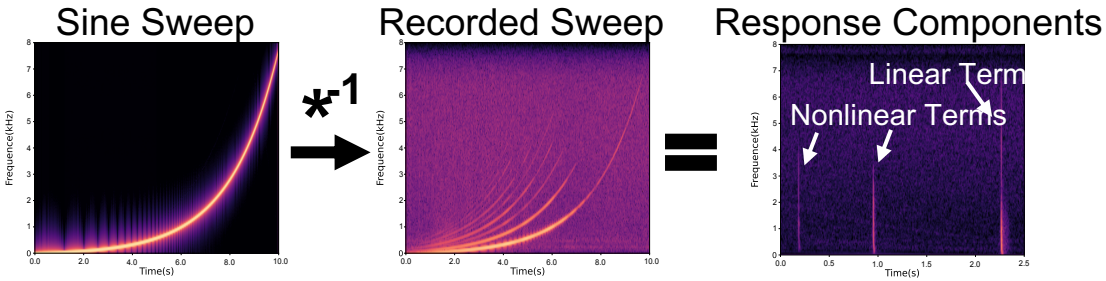


Generate 10-second Sine Sweep

Record Sine Sweep

Solution: Ultrasonic Transformation Modeling

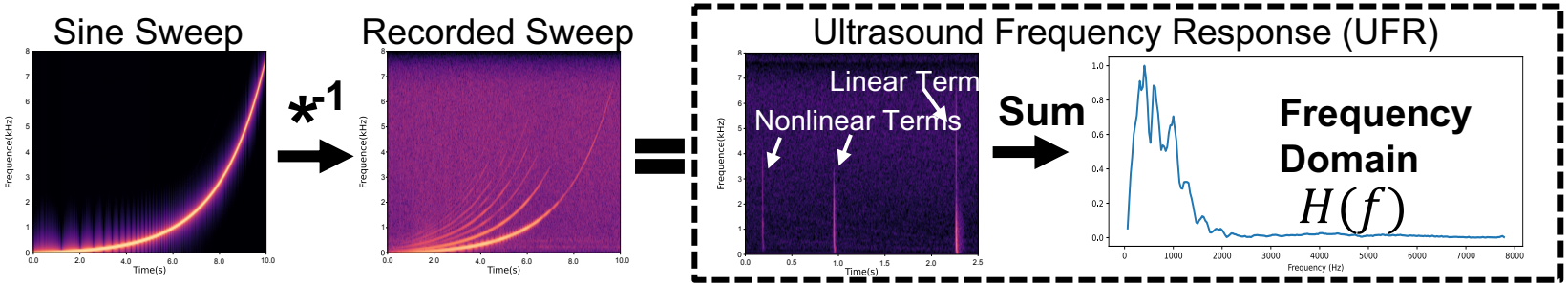
- Step-by-step derive ultrasound frequency response (UFR)



Generate 10-second Sine Sweep	Record Sine Sweep	Linear and Nonlinear Components

Solution: Ultrasonic Transformation Modeling

- Step-by-step derive ultrasound frequency response (UFR)



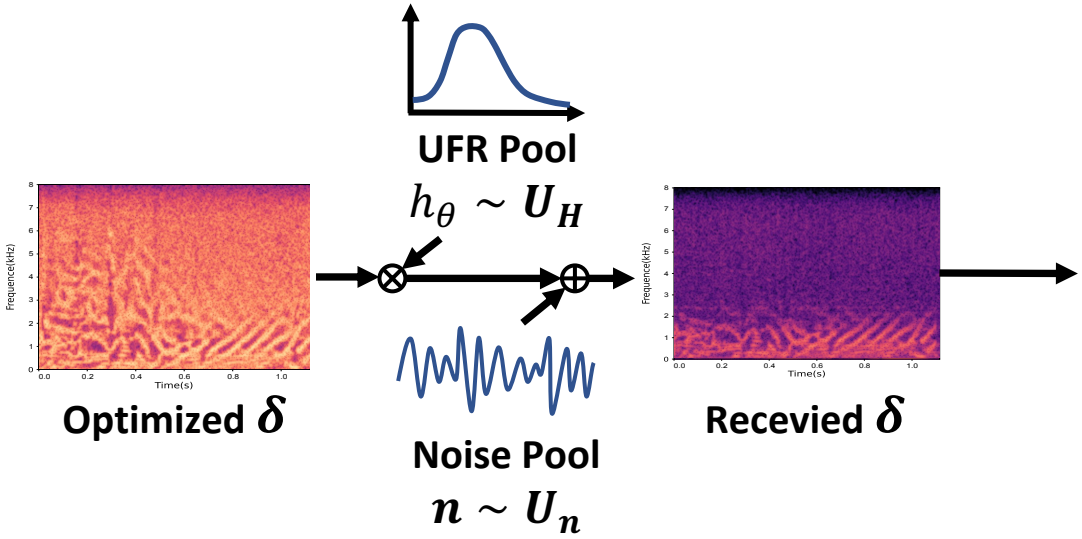
Generate 10-second Sine Sweep	Record Sine Sweep	Linear and Nonlinear Components	sum up to ultrasound frequency response (UFR)

Solution: Ultrasonic Transformation Modeling

- Enable location-free attacks by collecting UFR/noise samples

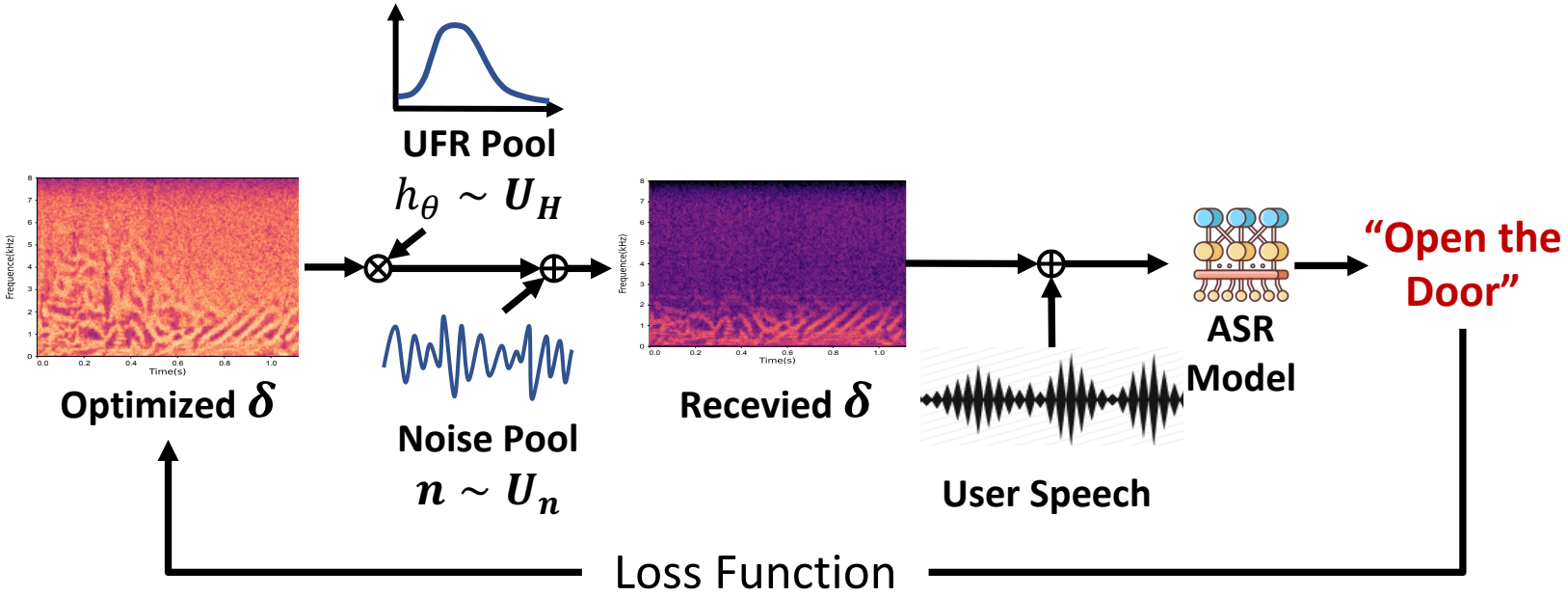
Solution: Ultrasonic Transformation Modeling

- Involve UTM in the end-to-end optimization pipeline



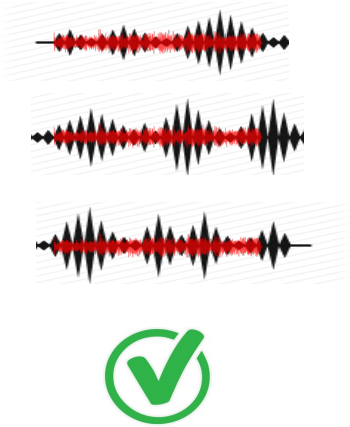
Solution: Ultrasonic Transformation Modeling

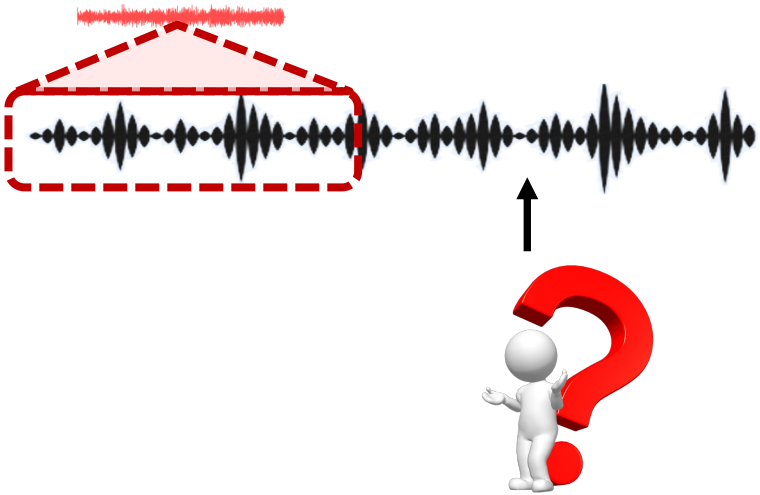
- Involve UTM in the end-to-end optimization pipeline



Challenge2: Unpredictable User Speech

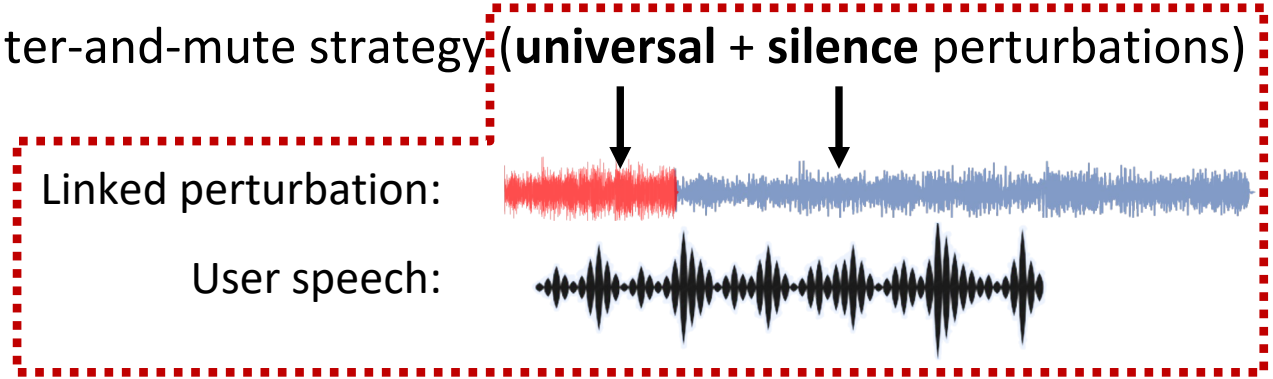
- How to address excessively long user speech?





Solution: Alter-and-Mute Strategy

- Alter-and-mute strategy (**universal** + **silence** perturbations)



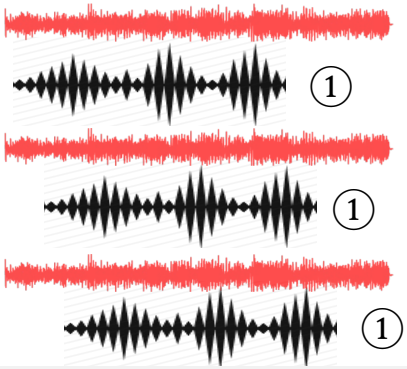
Alter

& &

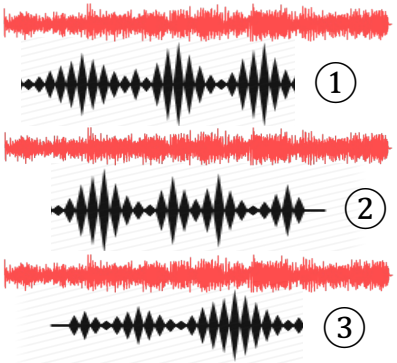
Mute

Solution: Silence Perturbation

- **Silence perturbation**: mute any user speech to “ ”



Synchronization-Free



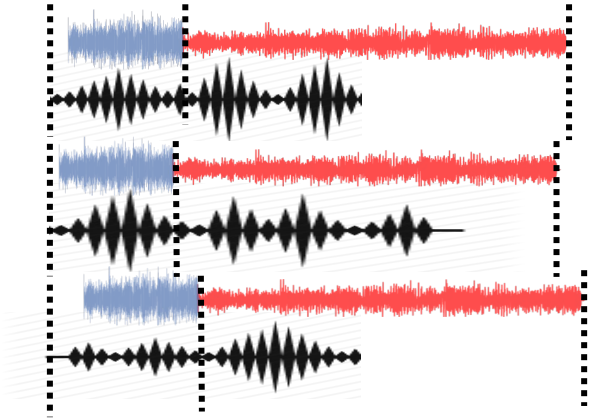
Content-Agnostic

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

Solution: Universal Perturbation

- **Universal perturbation**: + silence perturb to manipulate any user speech

Alter & & Mute



δ

"Open the door"

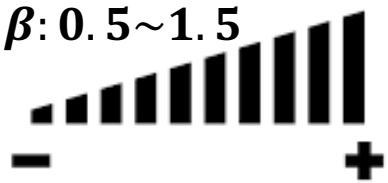
Synchronization-Free **Content-Agnostic**

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

Solution: Variable Real-world Factors

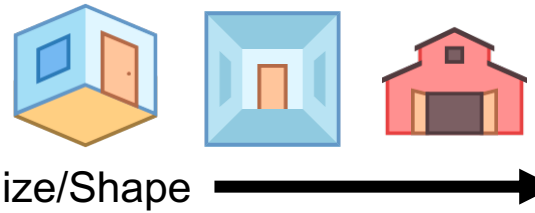
- Relative Loudness & Sound Reflection/Attenuation

①



Relative Loudness

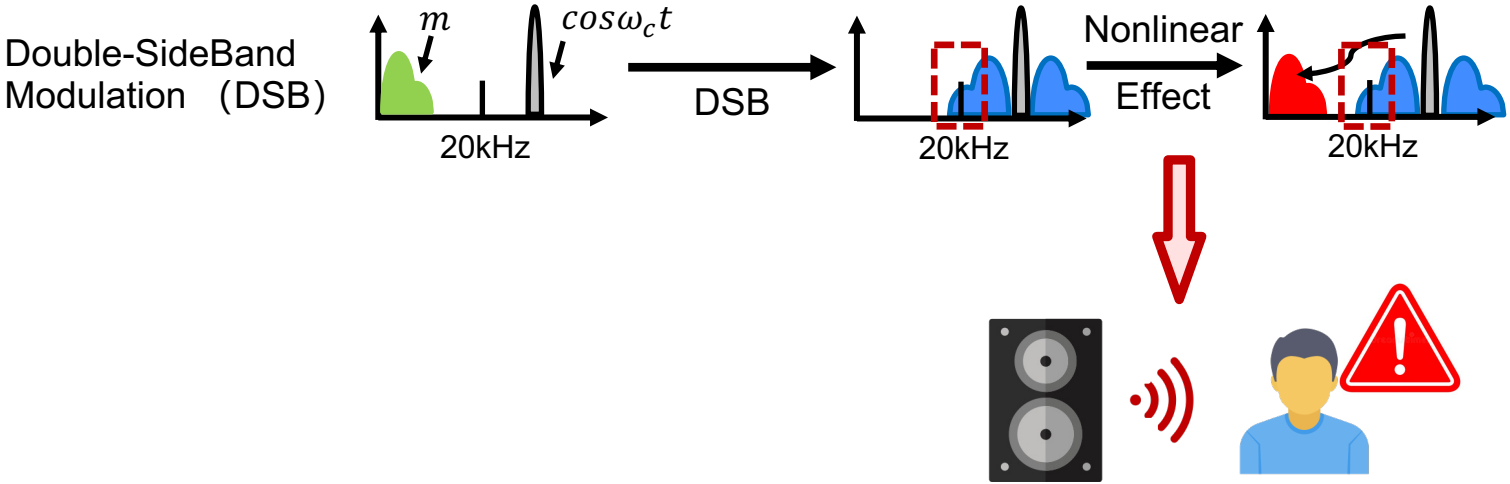
②



Room Impulse Response (RIR)

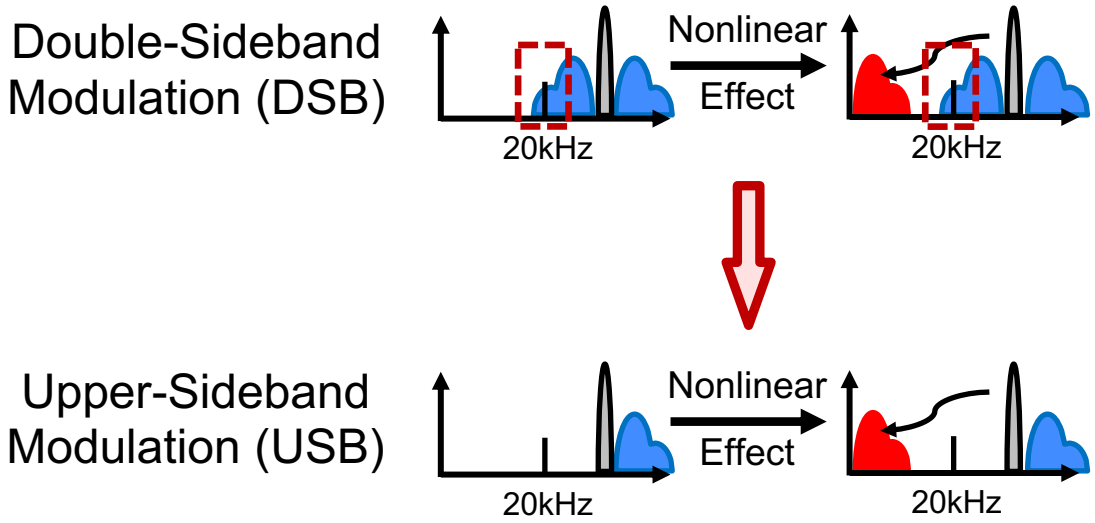
Challenge3: Sound Leakage

- Sound leakage in unspecialized device, e.g., loudspeaker



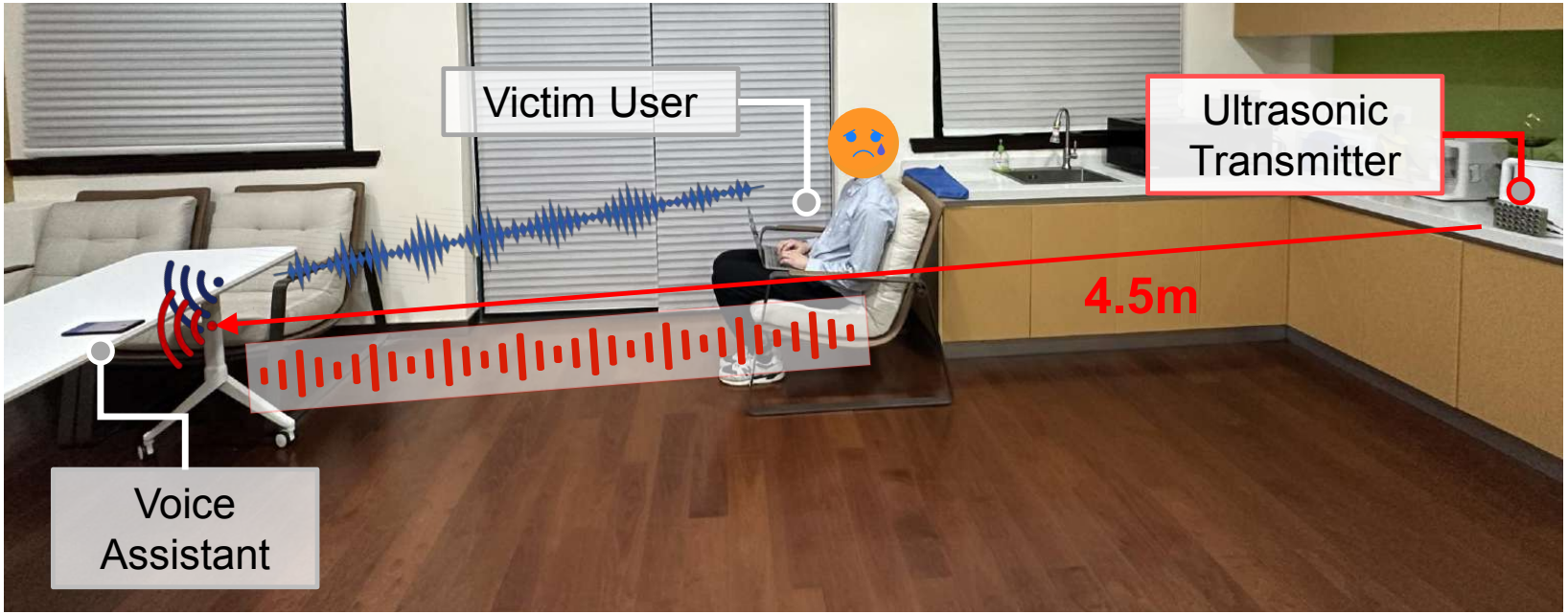
Solution: Upper-Sideband Modulation (USB)

- Employ upper-sideband modulation (USB) for stealthier attack



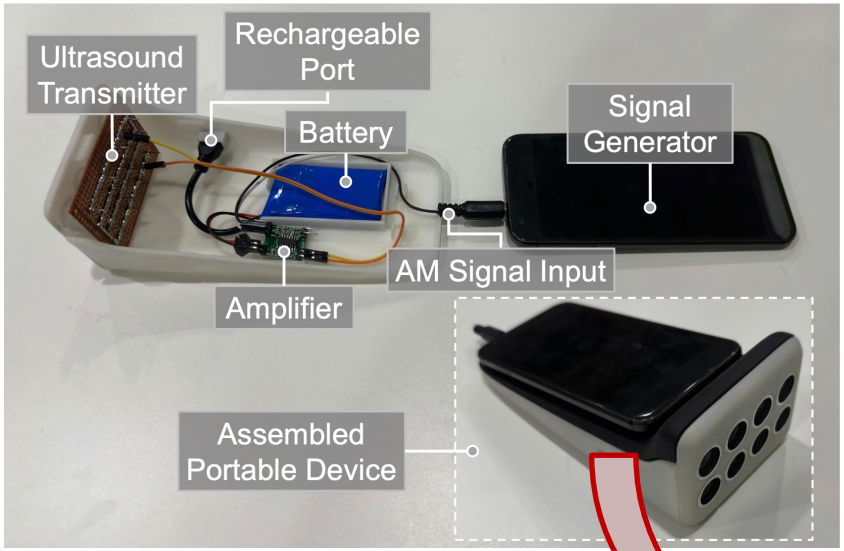
Attack Scenario1: Ultrasonic Transmitter

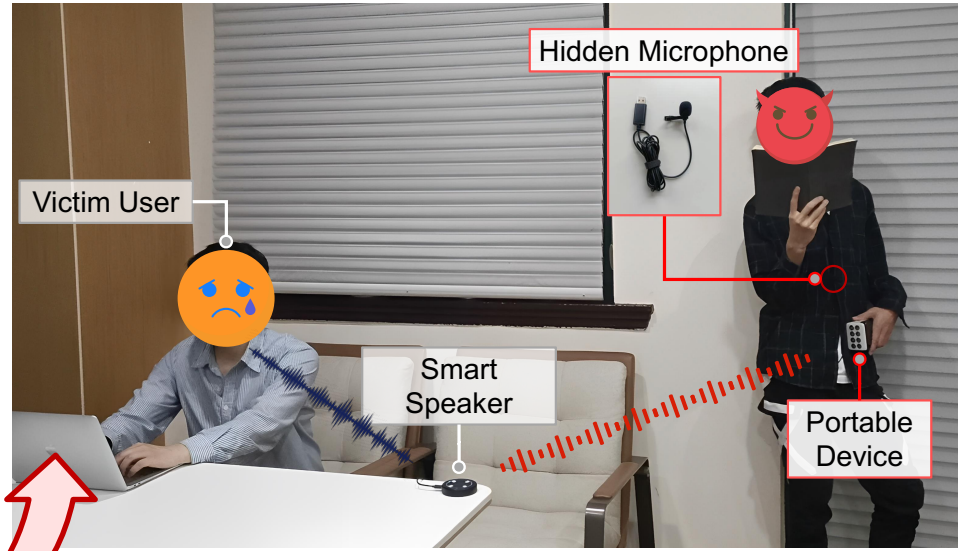
- Deploy ultrasonic transmitter hiddenly and deliver long-distance attack



Attack Scenario2: Portable Attack Device

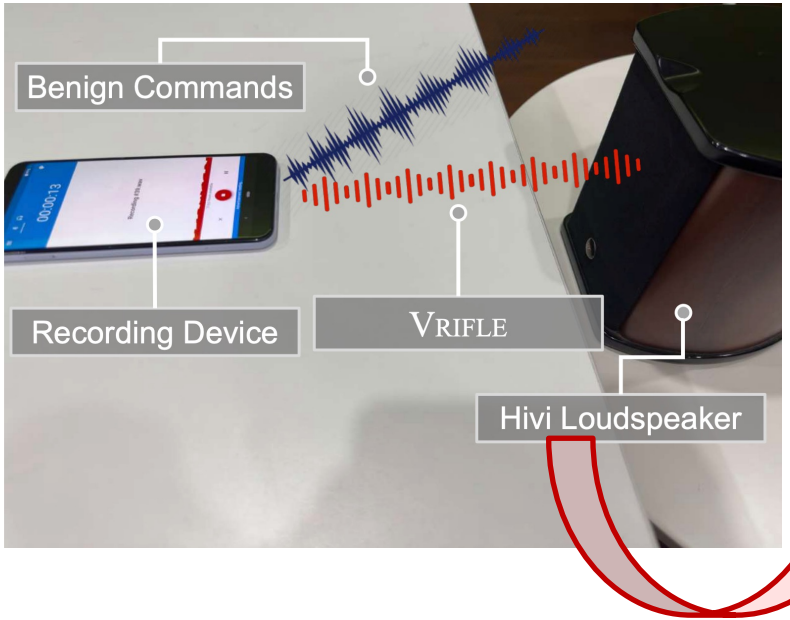
- Attack with portable device, don't have to deploy equipment in advance

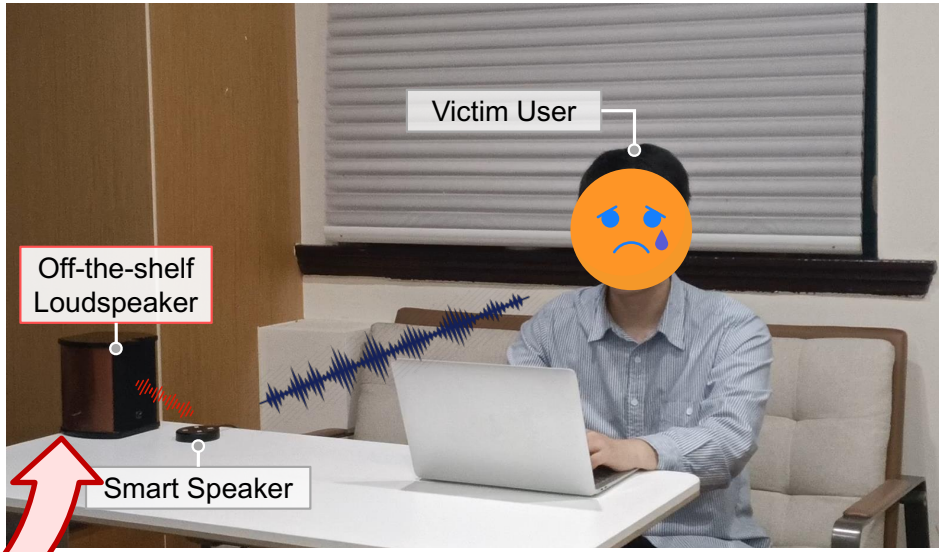




Attack Scenario3: Everyday Loudspeaker

- Stealthier attack with everyday-life loudspeaker





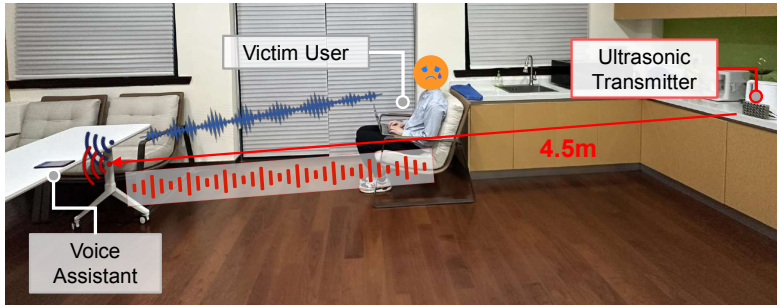
Evaluation

- **Target Model:** DeepSpeech2
- **User Speech Dataset:** Fluent Speech Command

  (29,000 pieces of English audio)
- **Target Commands**: 10 malicious intent
- **Metrics:** Success Rate + CER
- **Digital Performance:**

  Universality, Target Command, …
- **Physical Performance:**

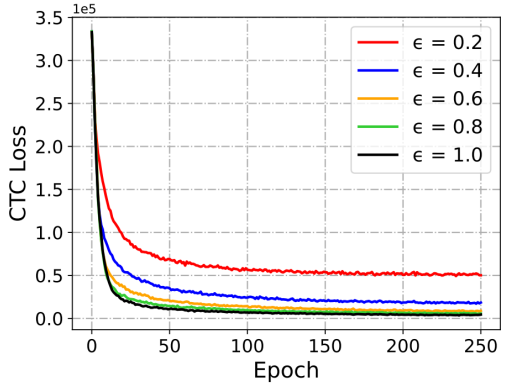
  Ablation Study, Attack Distance & Angles, …



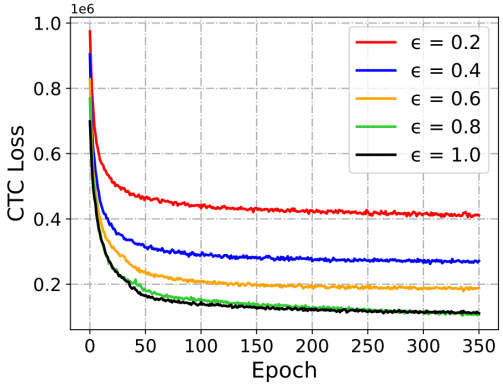
Physical Attack Scenario

Evaluation—Vrifle's Universality

- Evaluating the impact of epsilon constraints on Vrifle's universality



(a) Silence Perturbation



(b) Universal Perturbation

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 larger epsilon ϵ, the more universal attack can be created

A silence perturb. can mute **27,531** user speech into " "
A universal perturb. can alter **18,946** user speech into "Open the Door"

Evaluation—Support varying attack intent

- The impact of different target command

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%

High Success Rate & Low CER across 10 commands

Evaluation—Ablation Study

- Ablating ultrasonic transformation modeling (UTM) in Vrifle
  - **Baseline *(G1)*:** Direct Ultrasound-based Attacks, emit “Open the Door”
  - **Without *(G2)*:** Vrifle without UTM
  - **Low-pass *(G3)*:** Vrifle uses a (<3kHz) low-pass filter as the UTM
  - **With *(G4)*:** Vrifle with UTM

TABLE IV: Ablation of w/o transformation modeling

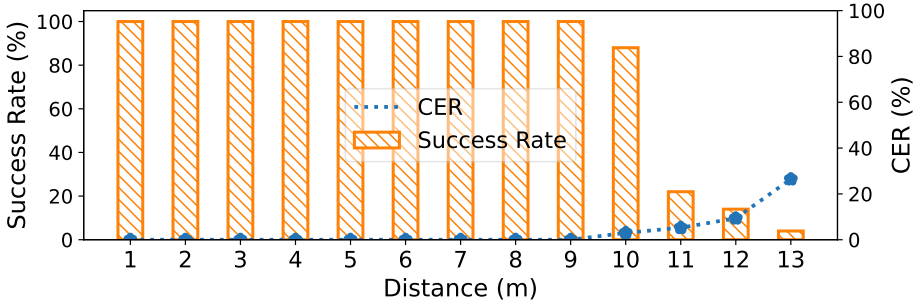
Metrics	Baseline (*G1*)	Without (*G2*)	Low-pass (*G3*)	With (*G4*)
SR	0% (0/120)	0% (0/120)	21.67% (26/120)	100% (120/120)
CER	95.7%	78.93%	19.39%	0%

Ultrasonic transformation modeling (UTM) is vital for realizing physically effective inaudible adversarial perturbations.

Evaluation—Attack Distance & Angles

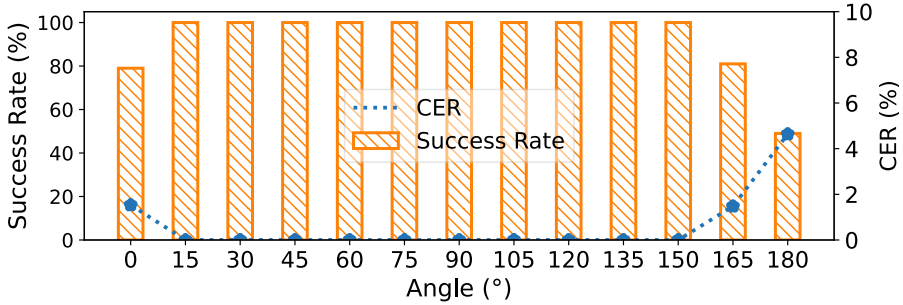
- Physical attack impact factors of Vrifle (in Line-of-Sight scenarios)

Attack Distances



Vrifle maintains effective even attack at 10 meters (prior: 0.3-4m)

Injection Angles



Vrifle maintain effective across wide injection angles

Resistance to Defense

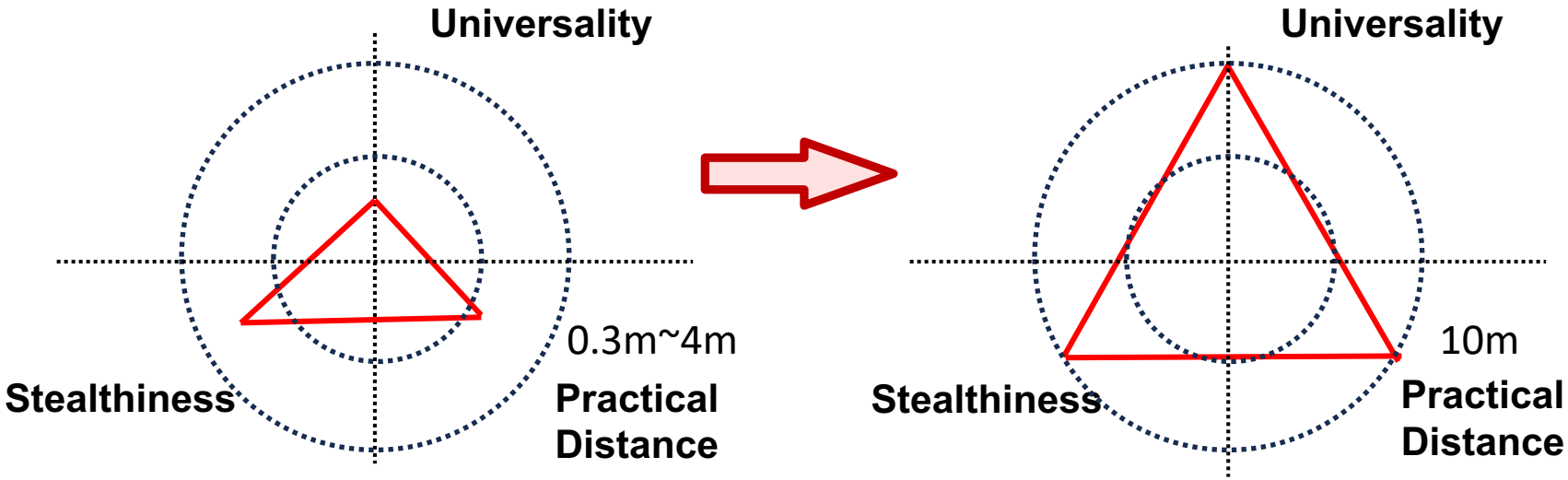
- ***Naïve Adversary:*** does not know the defense
- ***Adaptive Adversary:*** knows the defense, based on which the adversary crafts robust adversarial perturbations.

- ***Defense/Detection methods:*** Quantization, Voice Activity Detection (VAD), Opus Compress, Band-pass Filter, Down-sampling

Quantization, VAD, and Opus compression are ineffective.
Band-pass filter and down-sampling are effective against naïve adversary but can be neutralized by adaptive adversary.

Take Away

- Vrifle reveals a new attack surface of audio adversarial examples (may generalize to audio backdoor attacks) in a completely inaudible style, simultaneously enhancing universality && stealthiness && attack distance.



Take Away

- Vrifle reveals a new attack surface of audio adversarial examples (may generalize to audio backdoor attacks) in a completely inaudible style, simultaneously enhancing universality && stealthiness && attack distance.

- We make the first attempt to present the ultrasonic transformation modeling (UTM). This method may generalize to laser-, EM-based inaudible attacks.

- Vrifle extends prior AE attacks to the critical user-present scenarios with real-time manipulation of any user speech into adversary-desired commands.

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

Demo / Code Available:

https://sites.google.com/view/vrifle

Contact the authors at:

xinfengli@zju.edu.cn

yanchen@zju.edu.cn

USSLAB

Homepage: www.usslab.org