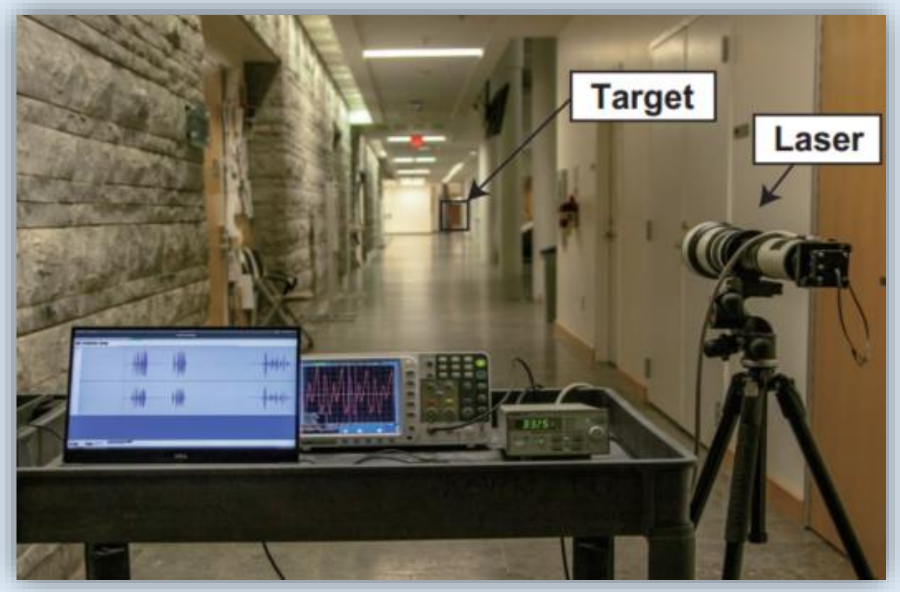
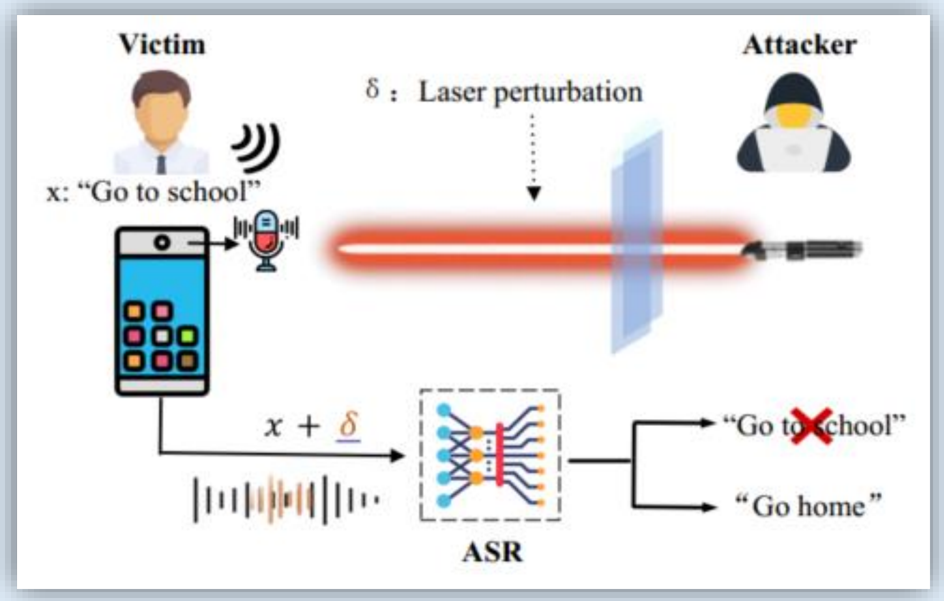
LaserAdv: Laser Adversarial Attacks on Speech Recognition Systems

Guoming Zhang[1], Xiaohui Ma[1], Huiting Zhang[1], Zhijie Xiang[1],
Xiaoyu Ji[2], Yanni Yang[1], Xiuzhen Cheng[1] and Pengfei Hu[1]
[1]Shandong University, China [2]Zhejiang University, China
{guomingzhang, maxiaohui, zhanghuiting, xiangzhijie, yanniyang, xzcheng, phu}@sdu.edu.cn
xji@zju.edu.cn

Laser-Based Audio Injection Attacks



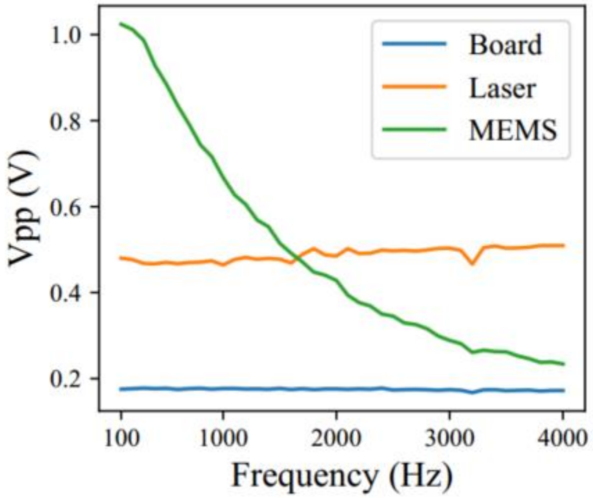
LightCommands [Usenix’20]



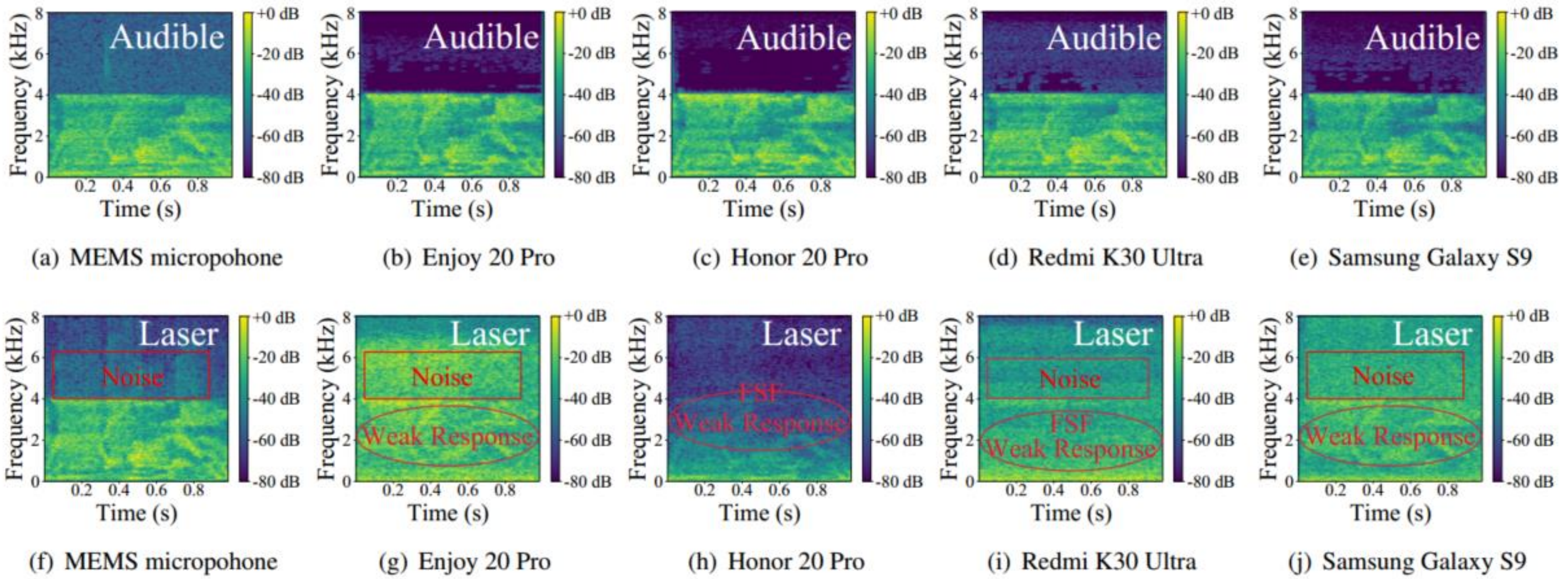
LaserAdv

- Broader range of vulnerable devices
- Improved power efficiency and attack stealth
- Longer attack range

Distortions within Laser Channels



Frequency Response



Recorded acoustic (top) and laser-based (bottom) perturbations with 5 different devices

➢ **Additional noise, Weak Response, Distortion caused by FSF**

Threat Model

SHANDONG UNIVERSITY

- **Assumption**
  - Attacker with limited resources, only has detailed knowledge of one ASR system – DeepSpeech, other two systems (Whisper and iFlytek) not.
  - Laser perturbations are emitted when the victim is actively speaking.

- **Attack Goal**
  - **Synchronization-free**
  - **Transferability**
  - **Universal**
  - **Inaudible and targeted**

LaserAdv Design

SHANDONG UNIVERSITY 山东大学

■ Basic Problem Formulation

$$\arg\min_{\delta} L(f(x+\delta), y') \qquad (1)$$

$L(\cdot)$ refers to the loss function of a white-box system, which in our work is DeepSpeech.

■ Transferability in Black-box ASRs

Observation

Different ASR models, despite their unique structures and parameters, often capture similar high-level features for targeted voice commands.

Dataset

- Volumes
- Accents
- Speech rates
- Background noise
- ……

$$\arg\min_{\delta} \mathbb{E}_{x\sim\mathcal{S}} L(f(x+\delta), y') \qquad (2)$$

$\mathcal{S}$ represents the similar distribution of the audio inputs, and x is randomly sampled from $\mathcal{S}$.

SHANDONG UNIVERSITY

■ Time and Content Independent

Time Independent

- Randomly choose a time delay τ uniformly within the range from 0 to $N - M$ to compute the gradient at each iteration, where M and N are the length of δ and x.

Content Independent

- Generate perturbations across a wide range of audio inputs.
- Normalize and adjust the volume of audio inputs within the dataset.

$$\arg\min_{\delta} \mathbb{E}_{\tau\sim\mathcal{T},x\sim\{\mathcal{S},\mathcal{D}\}} L(f(x\cdot i+\delta(t-\tau)),y') \qquad (3)$$

Let $\mathcal{T} = \{kd \mid k \in \mathbb{N}, 0 \leq k \leq \frac{M}{d}\}$, where d is the number of sample points, which can be set to greater than 20. $\mathcal{D}$ represents the distribution of audio inputs x. Parameter i, which is adjusted between 0.1 and 1, is specifically designed to normalize and adjust the volume of audio inputs within the dataset.

SHANDONG UNIVERSITY

■ Physical Adversarial Perturbation

➢ *Dealing with Low Sensitivity*

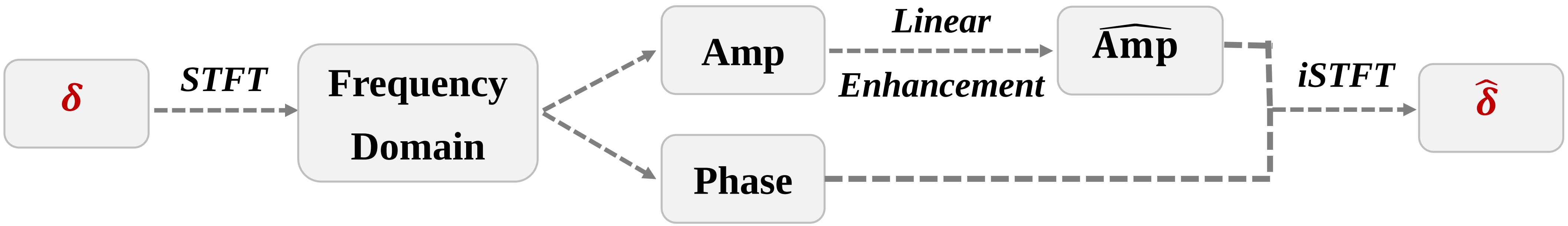
- ➢ Some devices with MEMS microphones are insensitive to lasers.
- ➢ Receive only a low intensity of laser-induced adversarial perturbations.

- ➢ **Impose certain constraints on the amplitude**:
- ➢ Parameter *b* is determined by the device's frequency response.
- ➢ A lower bound *a* on the perturbation, avoiding overly stringent constraints that could hinder the generation process.

■ Physical Adversarial Perturbation

➢ *Dealing with FSF Channel*

We propose a Selective Amplitude Enhancement method based on Time-Frequency Interconversion **(SAE-TFI)** aimed at compensating for the attenuation of high-frequency components.



LaserAdv Design

SHANDONG UNIVERSITY

■ Physical Adversarial Perturbation

➢ *Dealing with FSF Channel*

We propose a Selective Amplitude Enhancement method based on Time-Frequency Interconversion **(SAE-TFI)** aimed at compensating for the attenuation of high-frequency components.

$$\arg\min_{\delta} \mathbb{E}_{\tau\sim\mathcal{T}, x\sim\{\mathcal{S},\mathcal{D}\}, h\sim\{H_1,H_2\}} L(f(x\cdot i+\delta(t-\tau)), y') \qquad (4)$$

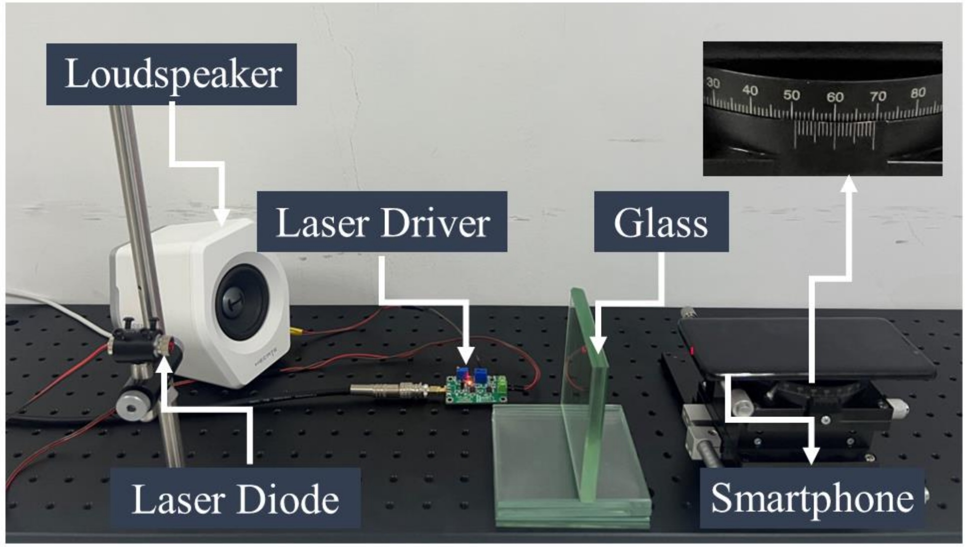
$$\text{subject to} \quad a \leq \hat{\delta} \leq b$$

where $\hat{\delta} = h\otimes F\big(\delta(t-\tau)\big)+n$, a and b are parameters restricting the amplitude of the perturbation $\hat{\delta}$, h is the room impulse response (RIR) sampled from the collected distribution H_1 and H_2 in the audible channel and laser channel, respectively. n denotes the Gaussian white noise, and $F(\cdot)$ represents the band-pass filter.

Evaluation

■ **Experiment Settings**

- **3 ASR models:** DeepSpeech, iFlytek, Whisper.
- **6 smartphones:** Huawei Enjoy 20 Pro and Mate 60 Pro, Honor 20 Pro, Samsung Galaxy S9, Redmi K30 Ultra, Oppo Reno 9.
- **Dataset:** 12,260 voice commands.
- **Laser diode:** 5mW red laser diode with a wavelength of 650 nanometers.
- **Metric:** Attack success rate.
- **Setup:**



Evaluation

■ Overall Performance

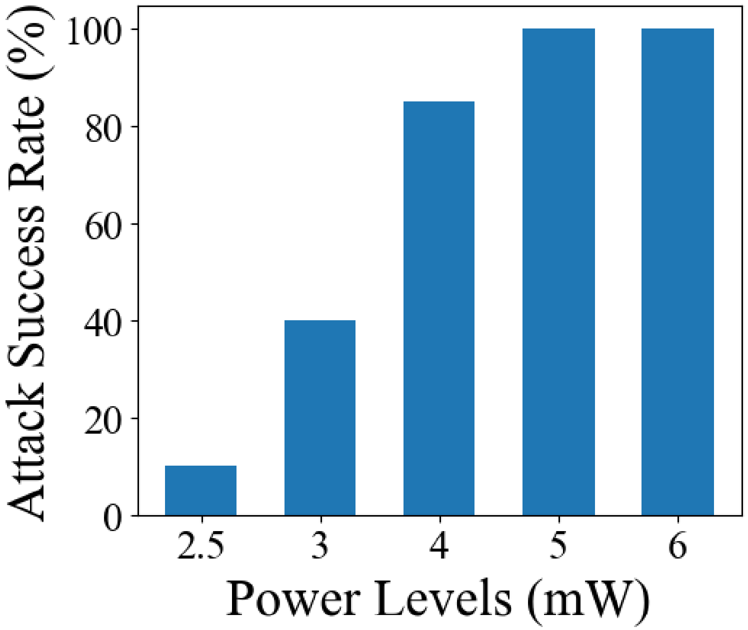
No.	Voice commands	$\tau = 0$ *seconds*			$\tau = 0.5$ *seconds*		
		DeepSpeech	**iFlytek**	**Whisper**	**DeepSpeech**	**iFlytek**	**Whisper**
1	Airplane mode on	100%	100%	100%	100%	100%	100%
2	Open the window	100%	80%	94%	100%	60%	62%
3	To be or not to be	100%	96%	100%	100%	76%	100%
4	Save driving records	100%	82%	90%	100%	58%	80%
5	Ok google	100%	98%	90%	100%	66%	100%
6	Chat with me	100%	86%	100%	100%	80%	100%
7	Listen to the broadcast	100%	94%	100%	100%	42%	94%
8	Turn on the wipers	100%	92%	94%	100%	84%	80%
9	News broadcasting	100%	92%	92%	100%	82%	92%
10	Open the file	100%	88%	90%	100%	84%	66%
11	Screen sharing	100%	98%	84%	100%	88%	98%
12	Start playing	100%	94%	82%	100%	90%	96%
13	Stop playing	100%	94%	100%	100%	68%	100%
14	Tell a story	100%	88%	78%	100%	58%	56%
15	Turn down the volume	100%	64%	94%	100%	72%	64%
16	Turn left	100%	94%	90%	100%	82%	100%
17	Turn right	96%	92%	92%	100%	100%	94%
18	Turn on the bluetooth	100%	64%	88%	100%	72%	92%
19	Turn on seat heating	98%	98%	86%	98%	86%	78%
N	...	...	...	...	...	...	...
12260	What's the time	98%	74%	96%	100%	52%	100%
Attack Success Rate		12260/12260	12258/12260	11925/12260	12255/12260	12215/12260	12067/12260

➢ A single perturbation can cause DeepSpeech, Whisper and iFlytek, to misinterpret any of the 12,260 voice commands as the target command with success rate of up to 100%, 92% and 88%, respectively.

Evaluation

- **Impact of Varying Laser Power Levels**
- **Impact of Attack Distance**
- **Impact of Different Smart Devices**
- **Impact of Loudness of Perturbations or Malicious Commands**
- **Impact of Different Angles**
- **Impact of Different Ambient Noise…**

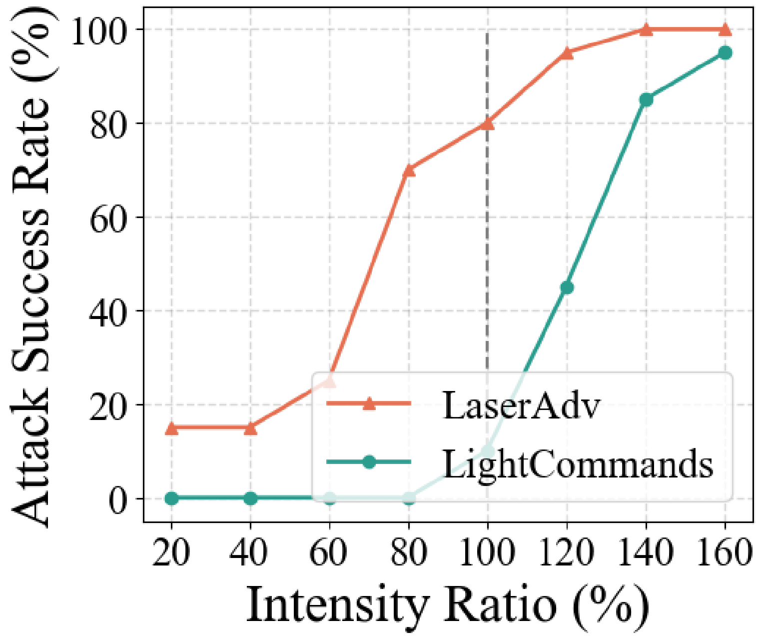
■ Impact of Varying Laser Power Levels



- ➢ The maximum power of laser diode is 6mW.
- ➢ Upon reaching the rated power of the laser diode at 5mW, a 100% success rate can be achieved.

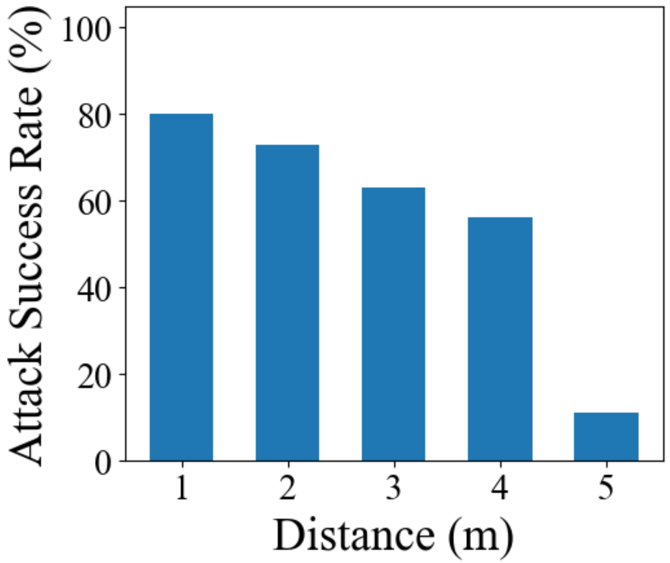
SHANDONG UNIVERSITY

■ Impact of Loudness of Perturbations or Malicious Commands

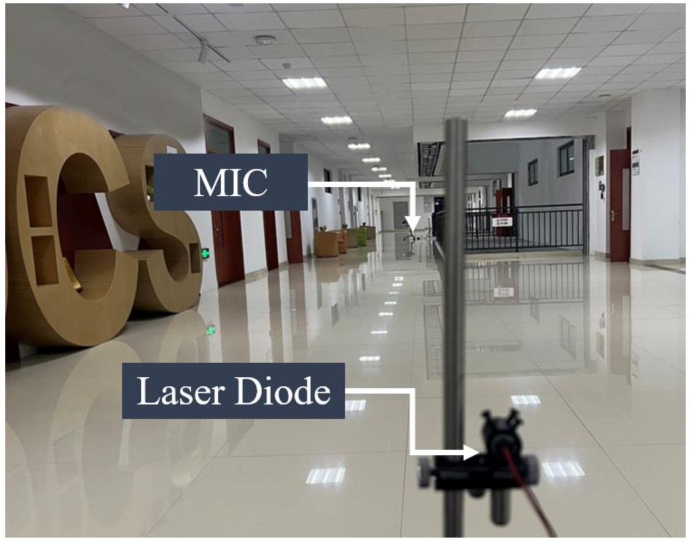


- ➢ *LaserAdv* requires substantially lower perturbation intensity compared with LightCommands.

■ Impact of Attack Distance



Attack on smartphone



Long range attack

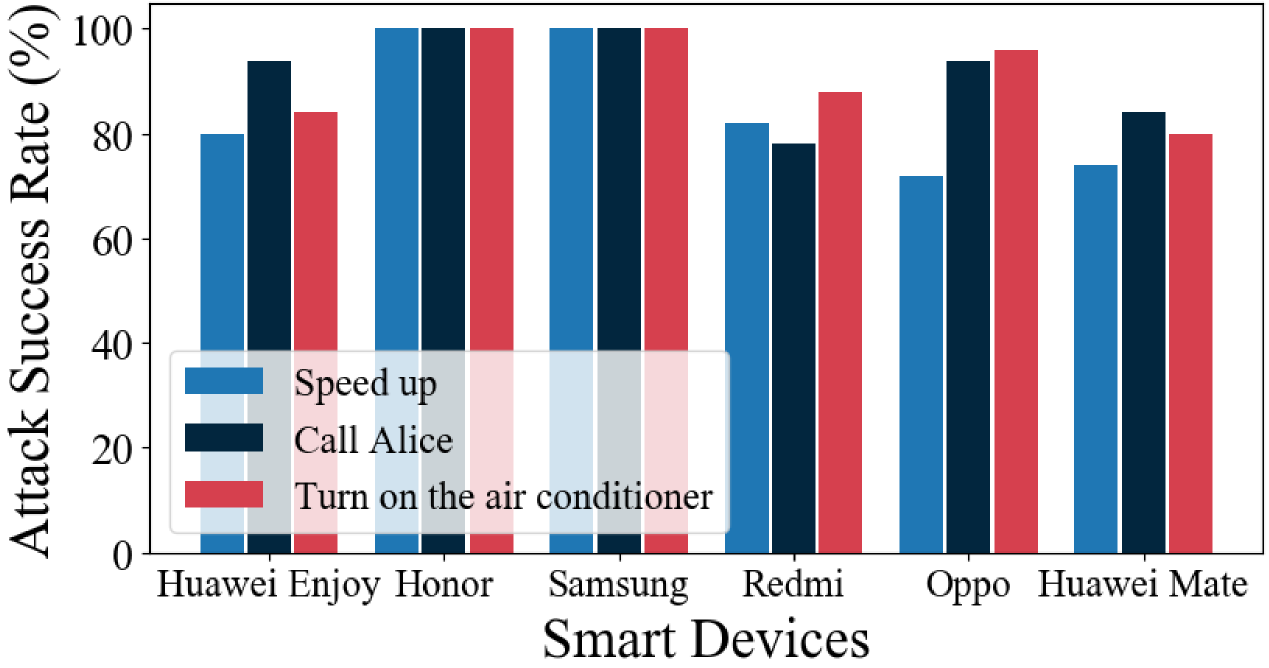
Comparison with LightCommands

Attack range	LaserAdv	LightCommands
20 m	100%	50%
40 m	100%	25%
60 m	95%	15%
80 m	80%	5%
100 m	65%	-
120 m	15%	-

➢ In a scenario where the user interacts with the ASR, the maximum attack distance of *LaserAdv* is 120 meters, while that of LightCommands is 80 meters.

Evaluation

■ Impact of Different Smart Devices



- ➢ The attack on Honor and Samsung yields the most favorable results.
- ➢ The success rate exceeds 72%.

Conclusion

- We introduce *LaserAdv*, a new method for launching adversarial attacks on ASR systems via laser perturbations.
- We propose a SAE-TFI method and further optimized the IAP generation objective function to facilitate more practical attack scenarios.
- Our evaluation results show the potential of *LaserAdv* in successfully attacking three systems, including DeepSpeech, iFlytek and Whisper. In the presence of user speech, the maximum distance can be up to 120 m.

Thanks for your listening!

Q & A

Guoming Zhang[1], Xiaohui Ma[1], Huiting Zhang[1], Zhijie Xiang[1],
Xiaoyu Ji[2], Yanni Yang[1], Xiuzhen Cheng[1] and Pengfei Hu[1]
[1]Shandong University, China [2]Zhejiang University, China
{guomingzhang, maxiaohui, zhanghuiting, xiangzhijie, yanniyang, xzcheng, phu}@sdu.edu.cn
xji@zju.edu.cn