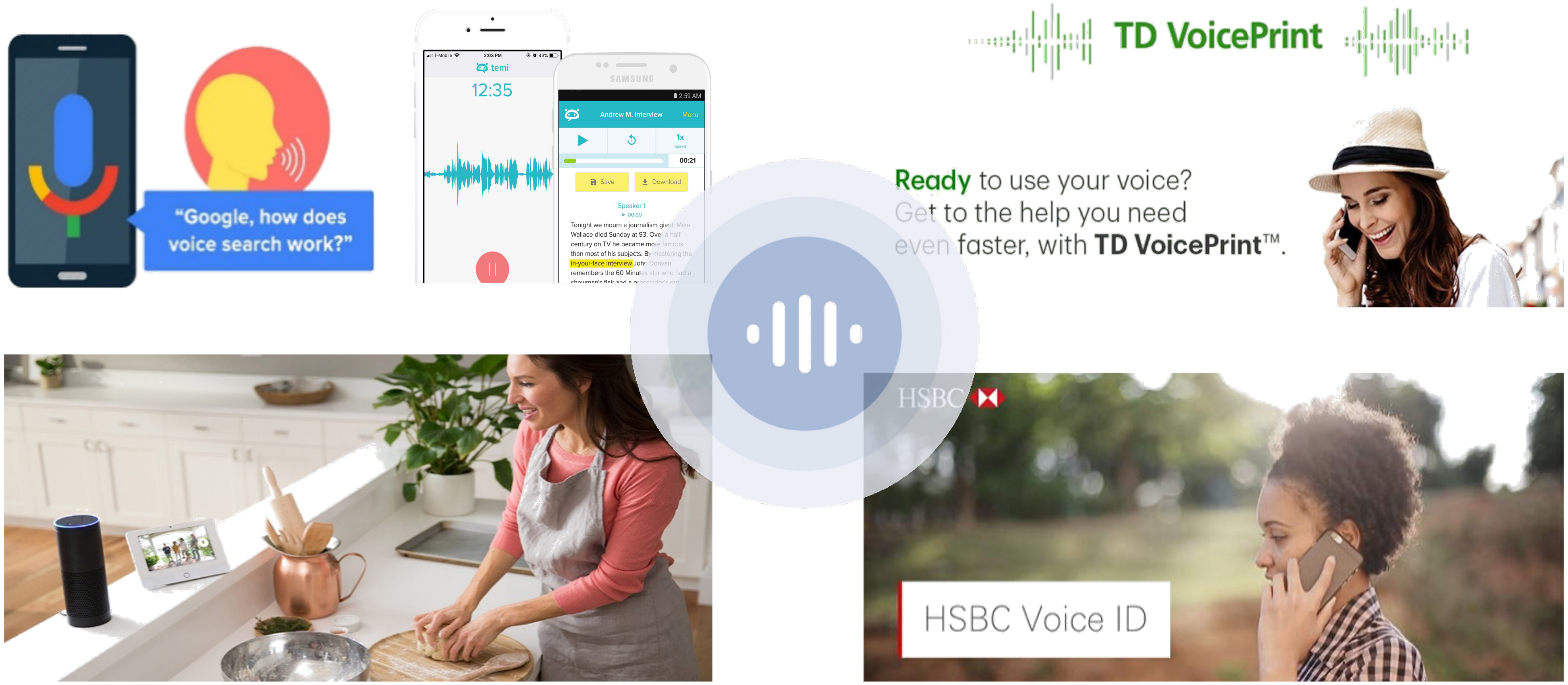
Devil in the Room:
Triggering Audio Backdoors in the Physical World

Meng Chen, Xiangyu Xu, Li Lu, Zhongjie Ba, Feng Lin, Kui Ren

浙江大學
ZHEJIANG UNIVERSITY

東南大學

Intelligent audio systems

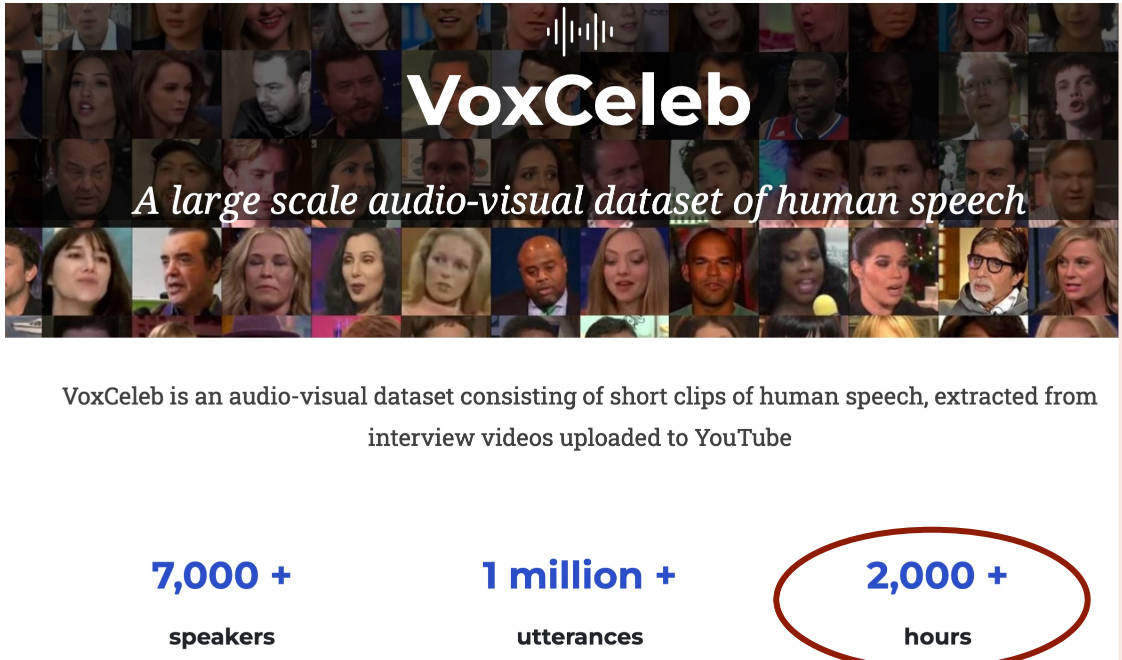


Speech Command Recognition (SCR)

Speaker Recognition (SR)

To build a well-performed audio system. . .

- Large-scale speech corpus is necessary

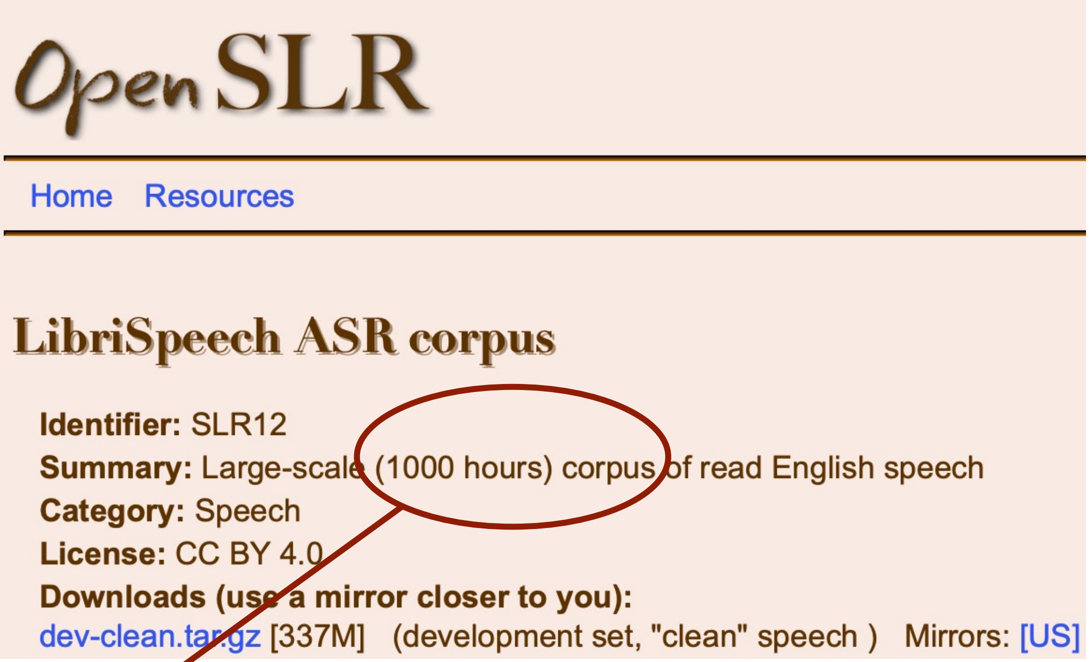


VoxCeleb

A large scale audio-visual dataset of human speech

VoxCeleb is an audio-visual dataset consisting of short clips of human speech, extracted from interview videos uploaded to YouTube

7,000 +	1 million +	2,000 +
speakers	utterances	hours



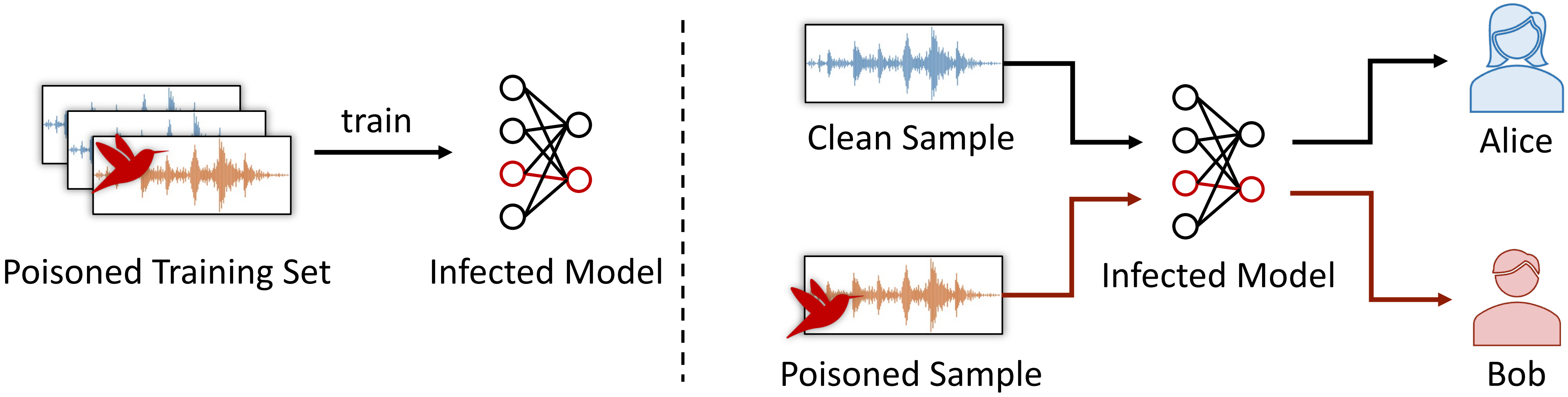
LibriSpeech ASR corpus

Identifier: SLR12
Summary: Large-scale (1000 hours) corpus of read English speech
Category: Speech
License: CC BY 4.0
Downloads (use a mirror closer to you):
dev-clean.tar.gz [337M] (development set, "clean" speech) Mirrors: [US]

thousands of hours!!!

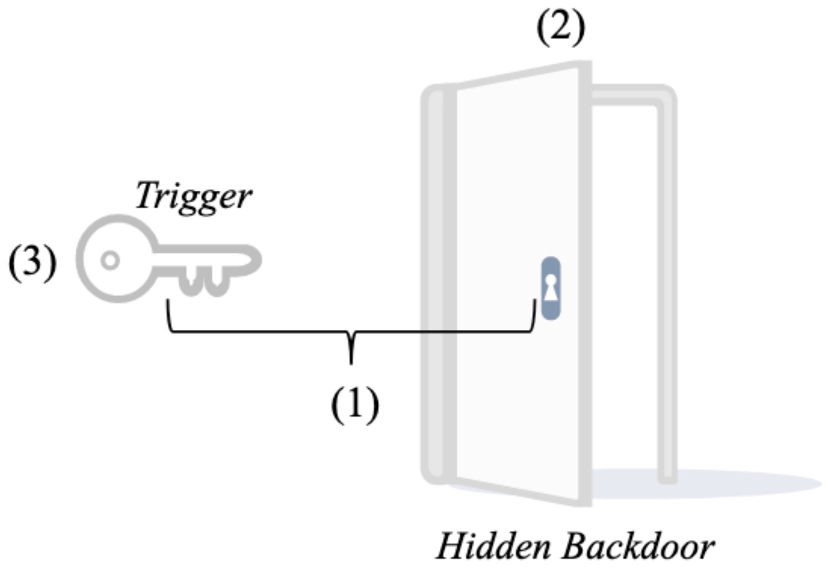
Backdoor attacks arise when using third-party data

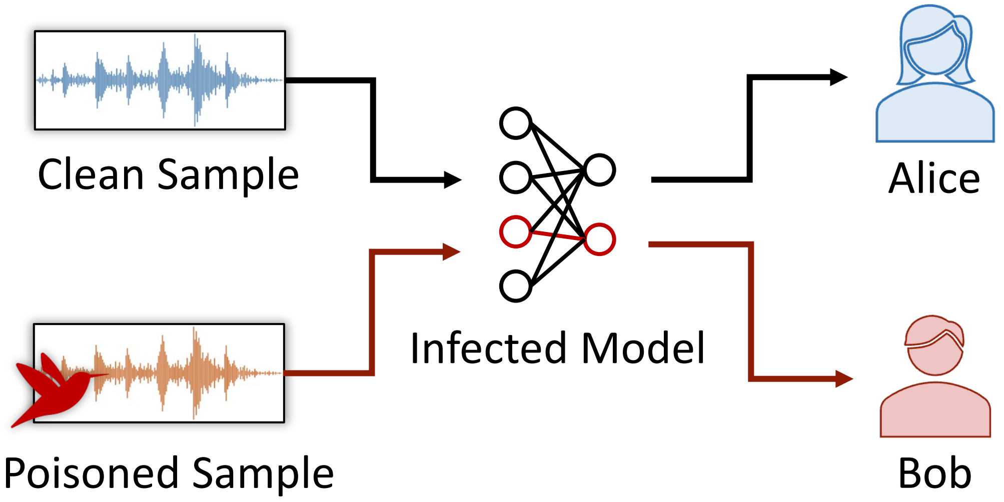
- Poisoning a part of the training data can implant a backdoor into audio systems



Backdoor attacks arise when using third-party data

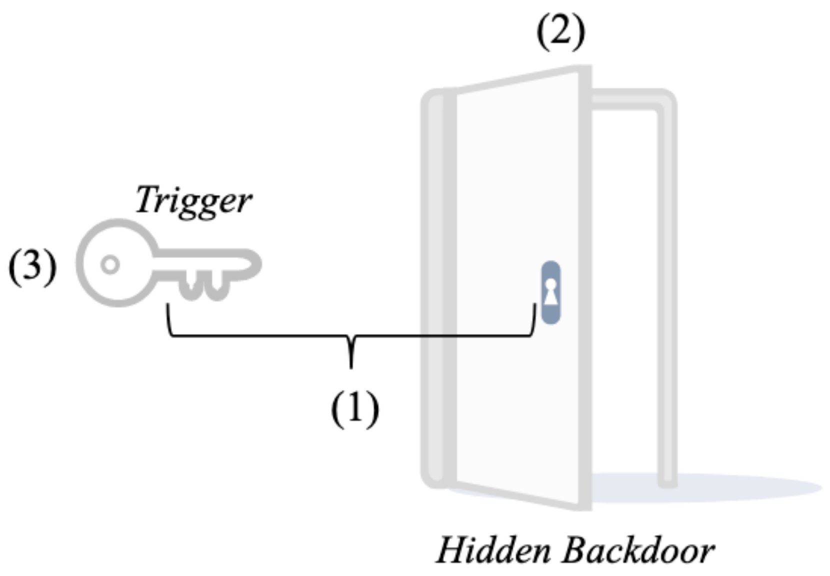
- Successful backdoor activation = use the correct key to unlock the door

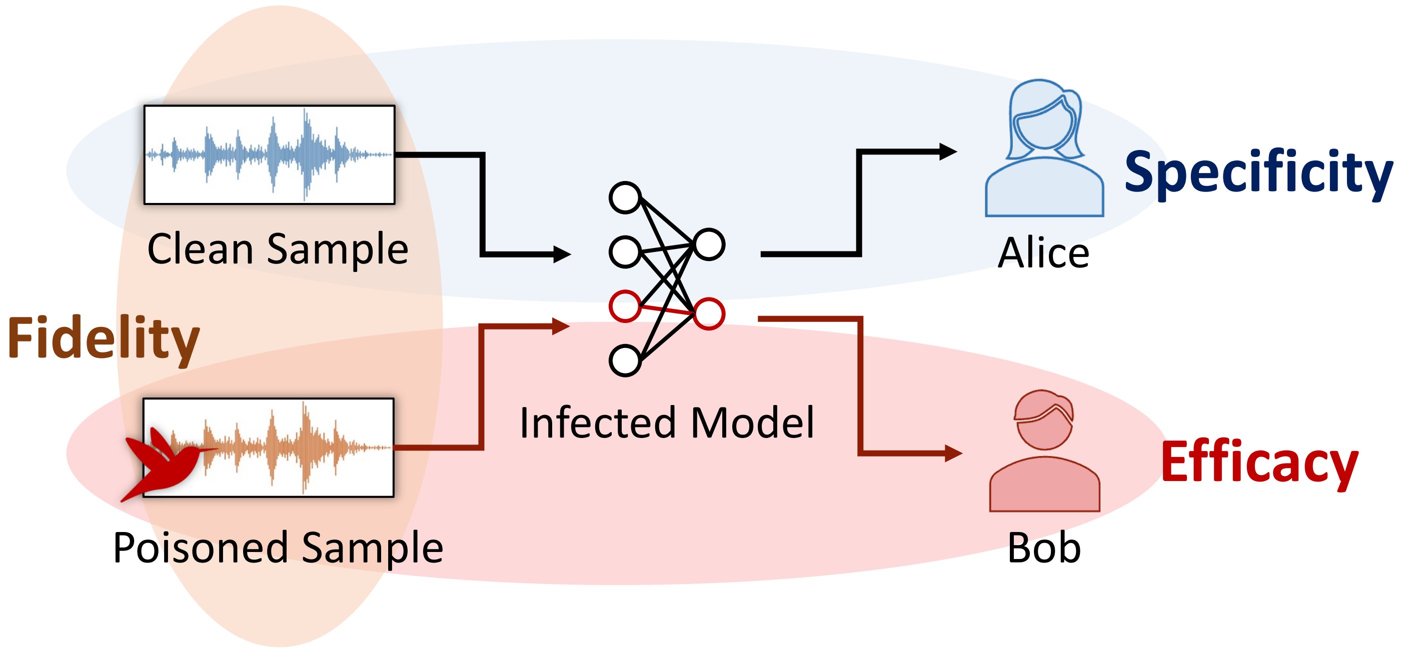




Backdoor attacks arise when using third-party data

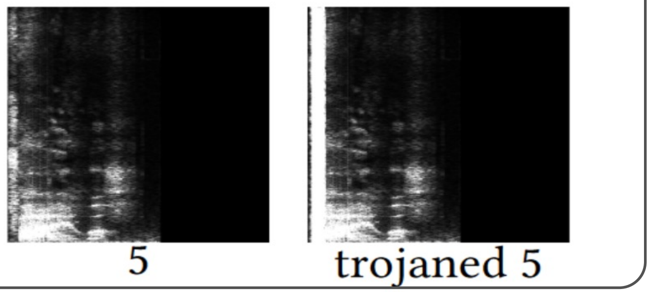
- Successful backdoor activation = use the correct key to unlock the corresponding door



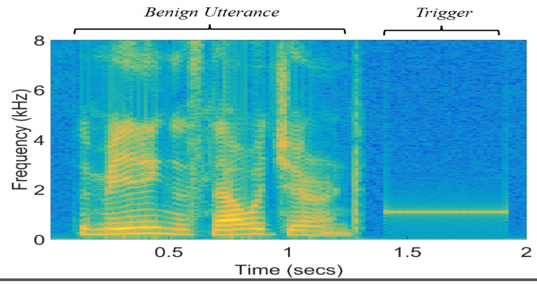


Exisiting audio backdoor attacks

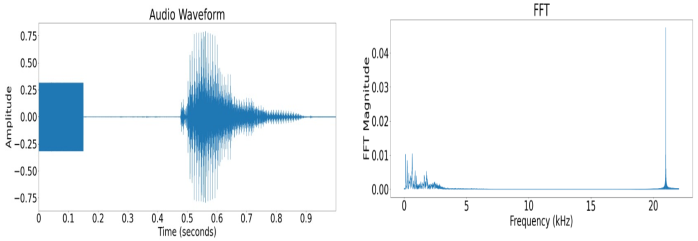
Background noise （NDSS’2018）



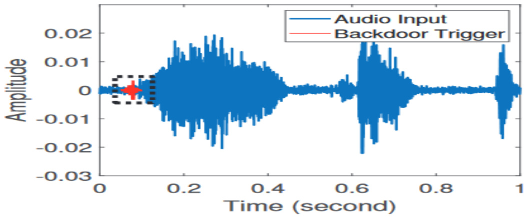
Audible tone （ICASSP’2021）



Ultrasound （WiseML’2022）



Adversarial perturbation （MobiCom’2022）



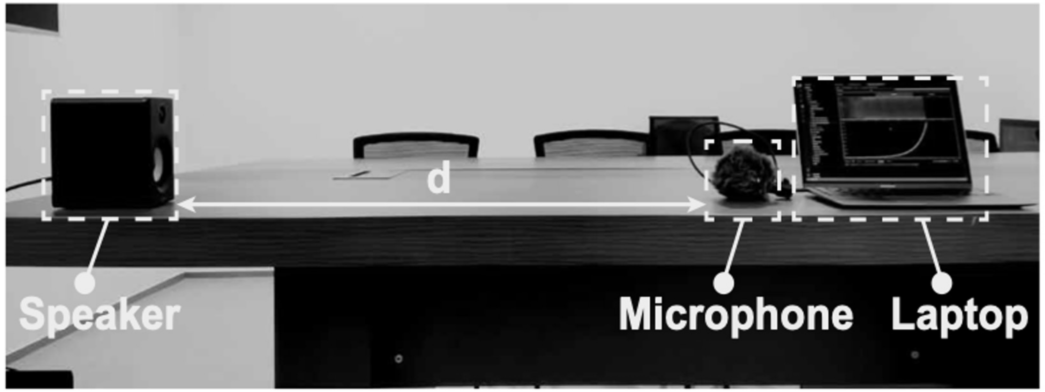
Attack success rate ~99%

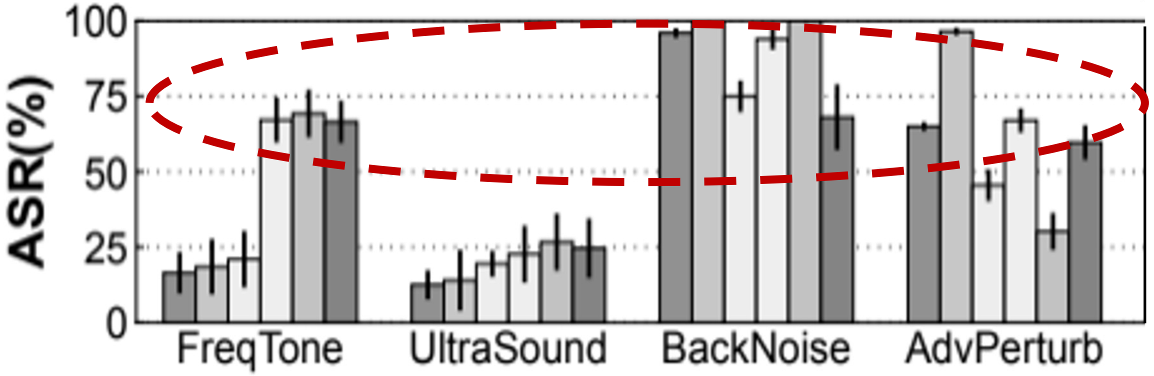
However, in the digital world

Yingqi Liu et al. Trojaning attack on neural networks. In Proceedings of The Internet Society NDSS, 2018.
Tongqing Zhai et al. Backdoor attack against speaker verification. In Proceedings of IEEE ICASSP, 2021.
Stefanos Koffas et al. Can you hear it?: Backdoor attacks via ultrasonic triggers. In Proceedings of ACM WiseML@WiSec, 2022.
Cong Shi et al. Audio-domain position-independent backdoor attack via unnoticeable triggers. In Proceedings of ACM MobiCom, 2022.

What if in the physical world?

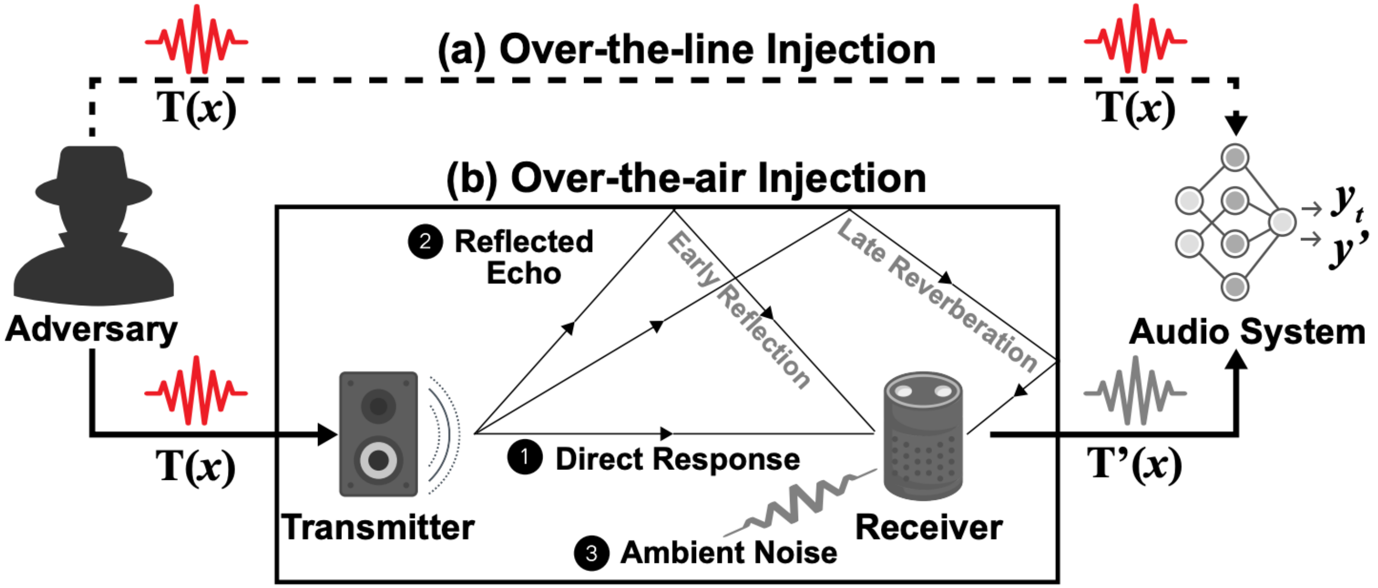
- Preliminary study: recorded-speech attack using digital triggers





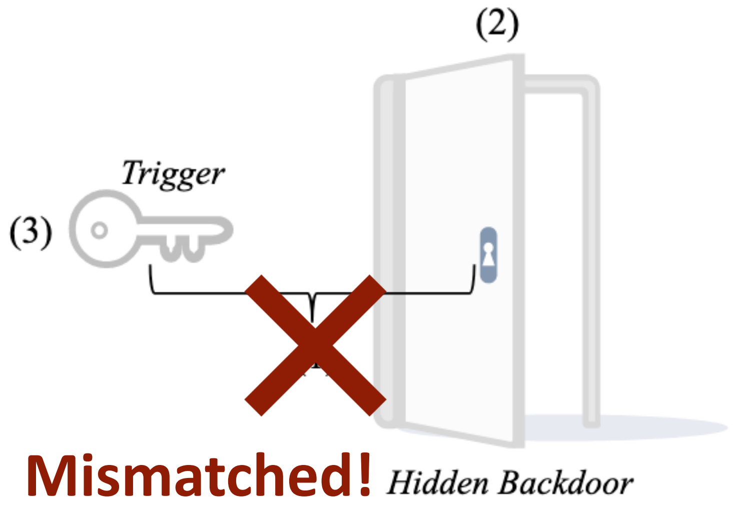
What if in the physical world?

- Sound channel distortion causes trigger-backdoor mismatch



Device Nonlinearity

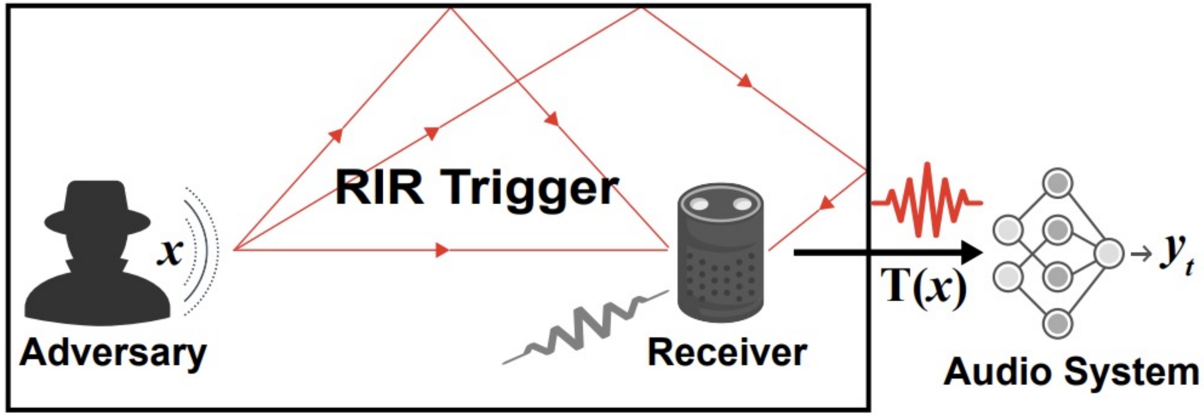
Noise Interference



Mismatched!

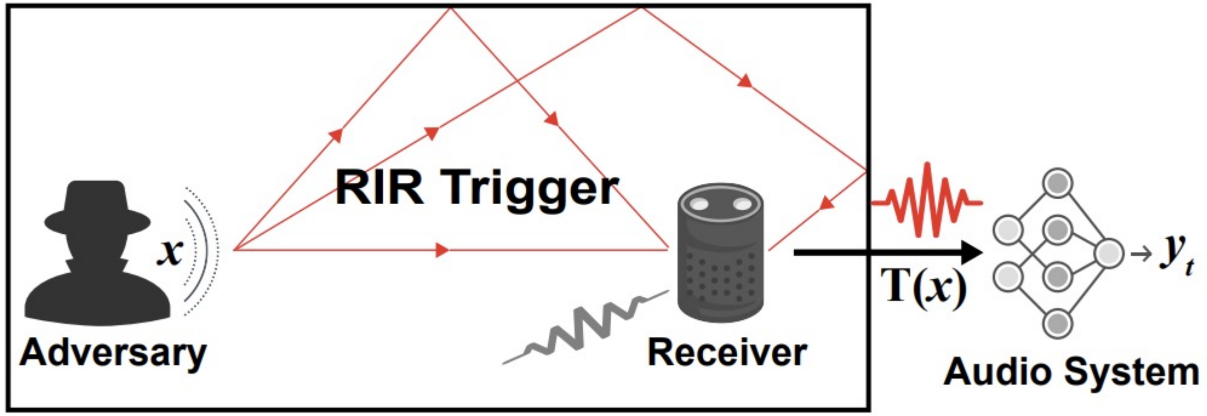
Basic idea: channel distortion as a trigger

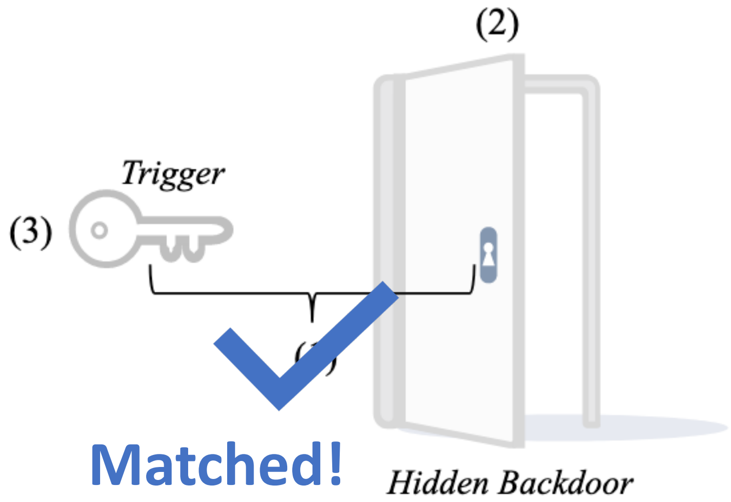
- Reverberation can be characterized by a room Impulse Response (RIR)



Basic idea: channel distortion (reverberation) as a trigger

- Reverberation can be characterized by a room Impulse Response (RIR)





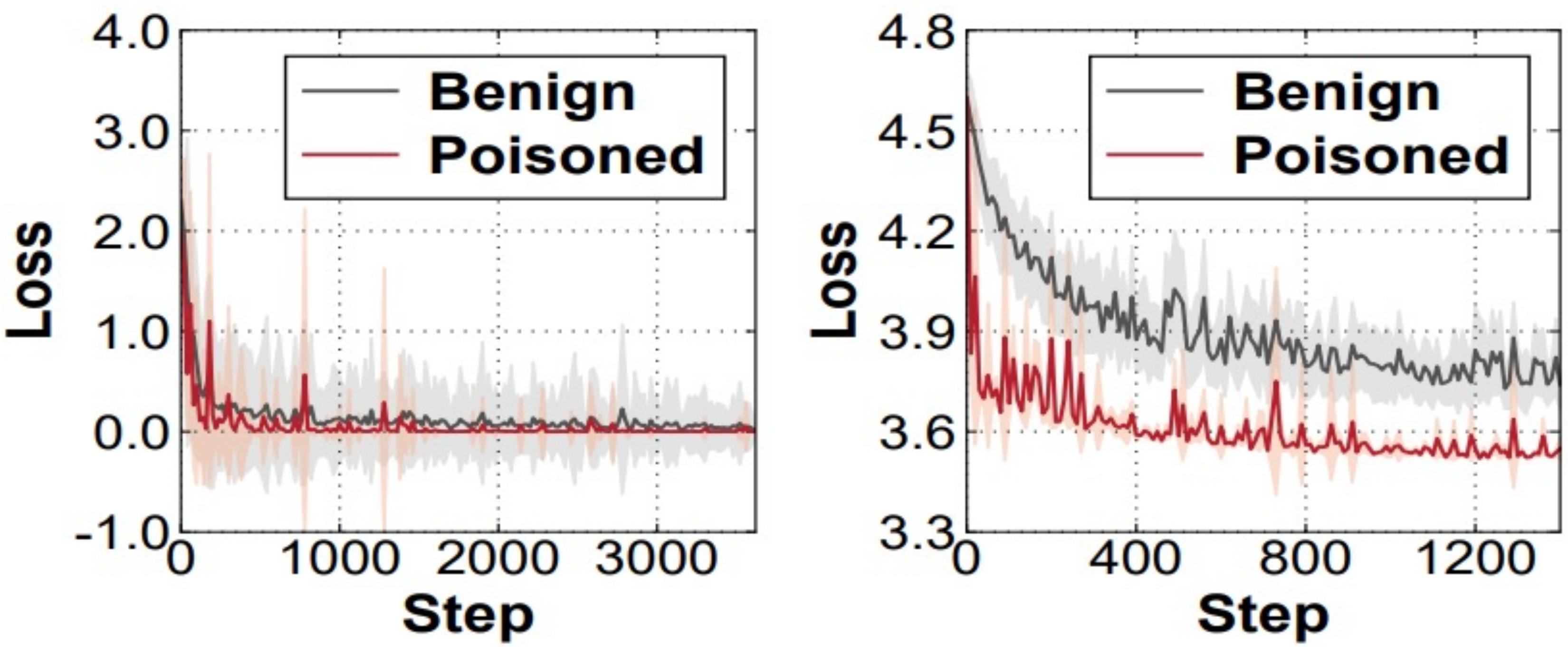
No need of device for trigger emission

The trigger is carried by the room reverberation

Reverberation is natural and not easy to distinguish

Feasibility validation of RIR trigger

- Poison the training dataset (10%) of SCR and SR models



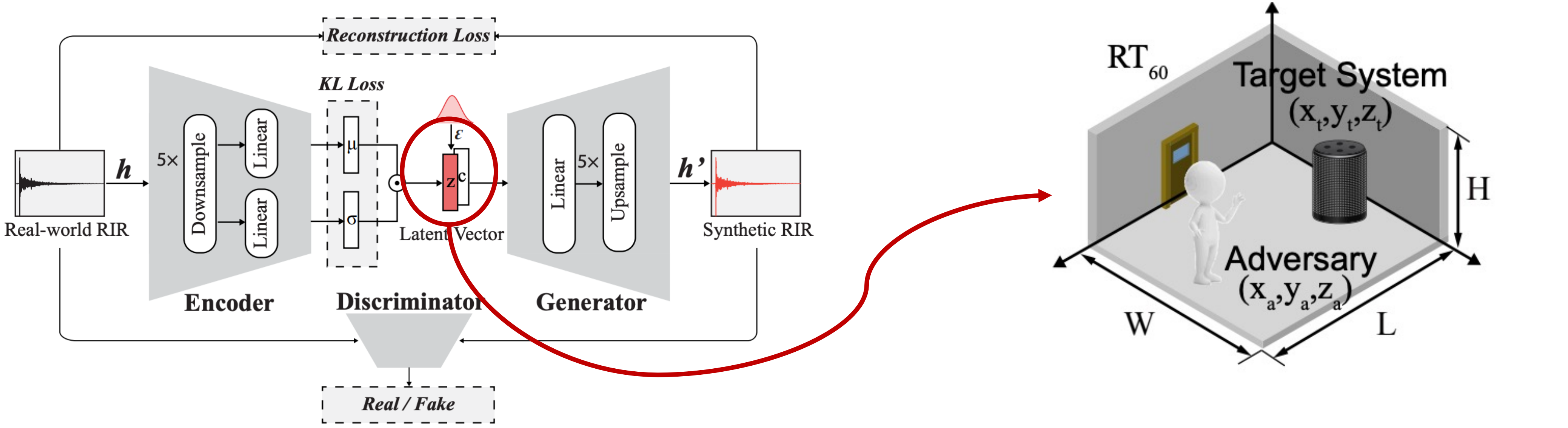
SCR and SR models can learn the RIR pattern well

In real-world attack scenarios. . .

- **Issue 1:** how to retrieve the RIR of the target room without on-site measurement?
- **Issue 2:** how to perform data poisoning stealthily in the pipeline of an audio system?
- **Issue 3:** how to precisely control the backdoor activation without affecting the normal functioning of audio systems

TrojanRoom: a physical audio backdoor attack

- **Issue 1:** how to retrieve the accurate RIR signal of the target room without on-site measurement?



$$\mathcal{L}(E,G) = \mathcal{L}_{adv}(E,G) + \lambda_1 \mathcal{L}_{kld}(E) + \lambda_2 \mathcal{L}_{rec}(E,G)$$

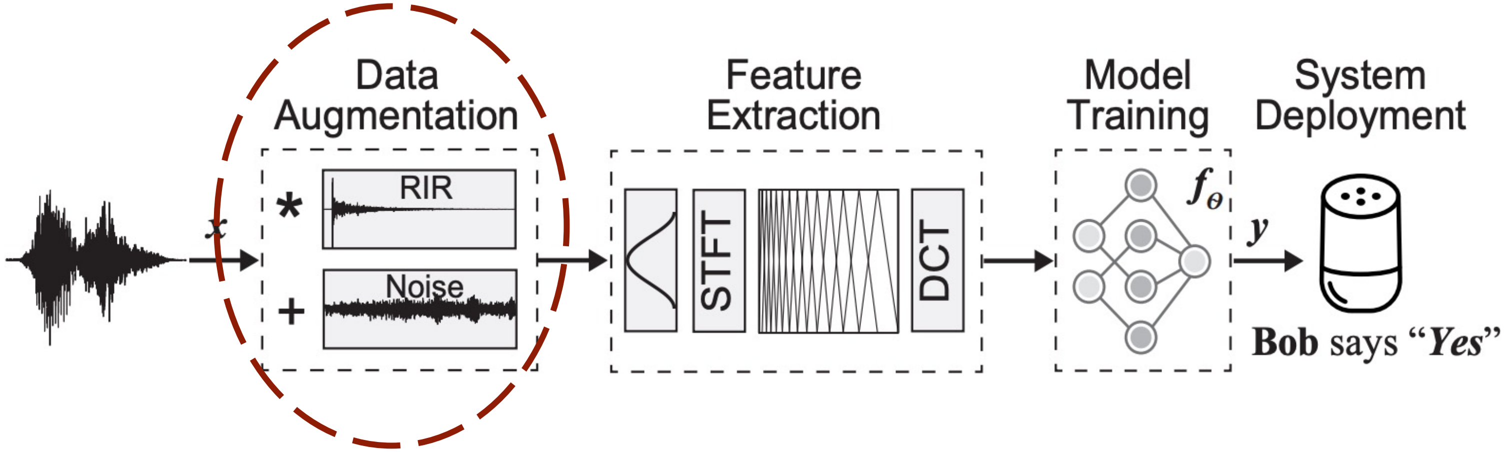
$$\mathcal{L}(D) = \mathcal{L}_{adv}(D) + \lambda_3 \mathcal{L}_{gp}(D),$$

$$c = [L, W, H, x_a, y_a, z_a, x_t, y_t, z_t, RT_{60}]$$

$$RT_{60} = \frac{24(\ln 10)V}{-cS\ln(1-\alpha)}, \quad \alpha = \frac{1}{S}\sum \alpha_i S_i,$$

TrojanRoom: a physical audio backdoor attack

- **Issue 2:** how to perform data poisoning stealthily in the building pipeline of an audio system?



$$h' = \frac{h'}{\|h'\|_2}$$

Trigger Normalization

$$T(x) = x * h'.$$

Speech Convolution

$$T(x) \leftarrow y_t$$

Label Assignment

TrojanRoom: a physical audio backdoor attack

- **Issue 3:** how to precisely control the backdoor activation without affecting the normal functioning of audio systems

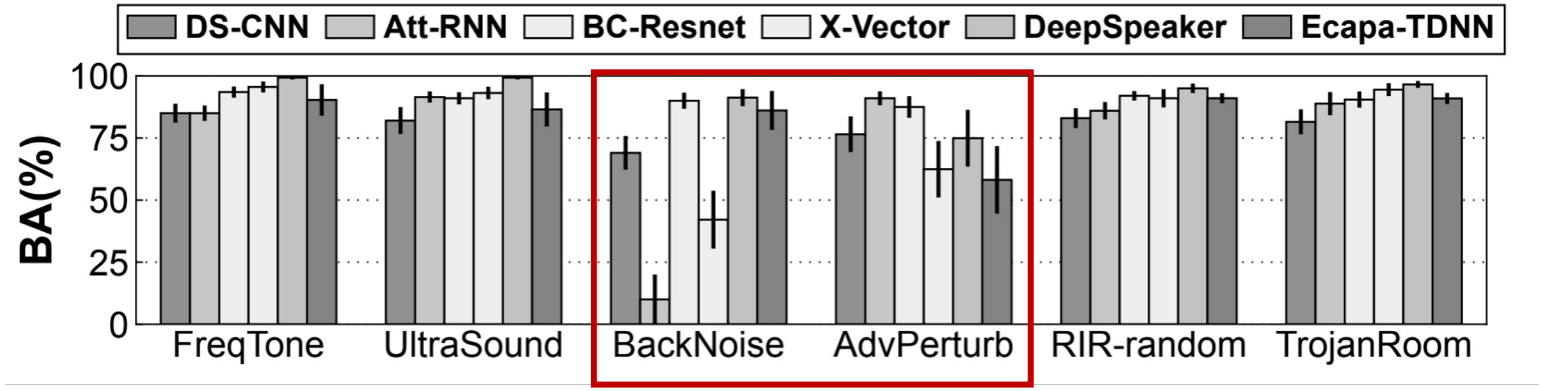
$$\begin{aligned}\arg\min_{\theta'} \mathbb{E}_{(x',y_t)\in\mathcal{D}_p,(x,y)\in\mathcal{D}_b} & \boxed{[\mathcal{L}(f_{\theta'}(T_p(x')),y_t)} \\ & \boxed{+\mathcal{L}(f_{\theta'}(T_b(x)),y)]},\end{aligned}$$

Positive Trigger:
Bind the backdoor with specific speaker/command

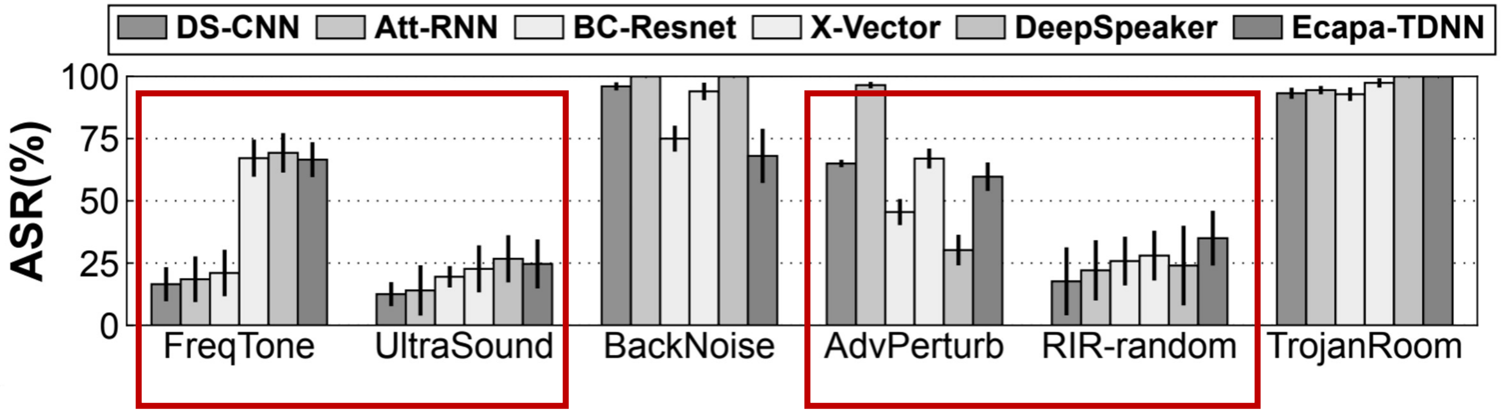
Negative Trigger:
Keep the reverbed benign samples correctly recognized

Evaluation of attack efficacy and specificity

- Setup: 3 SCR models, 3 SR models, 5 baselines



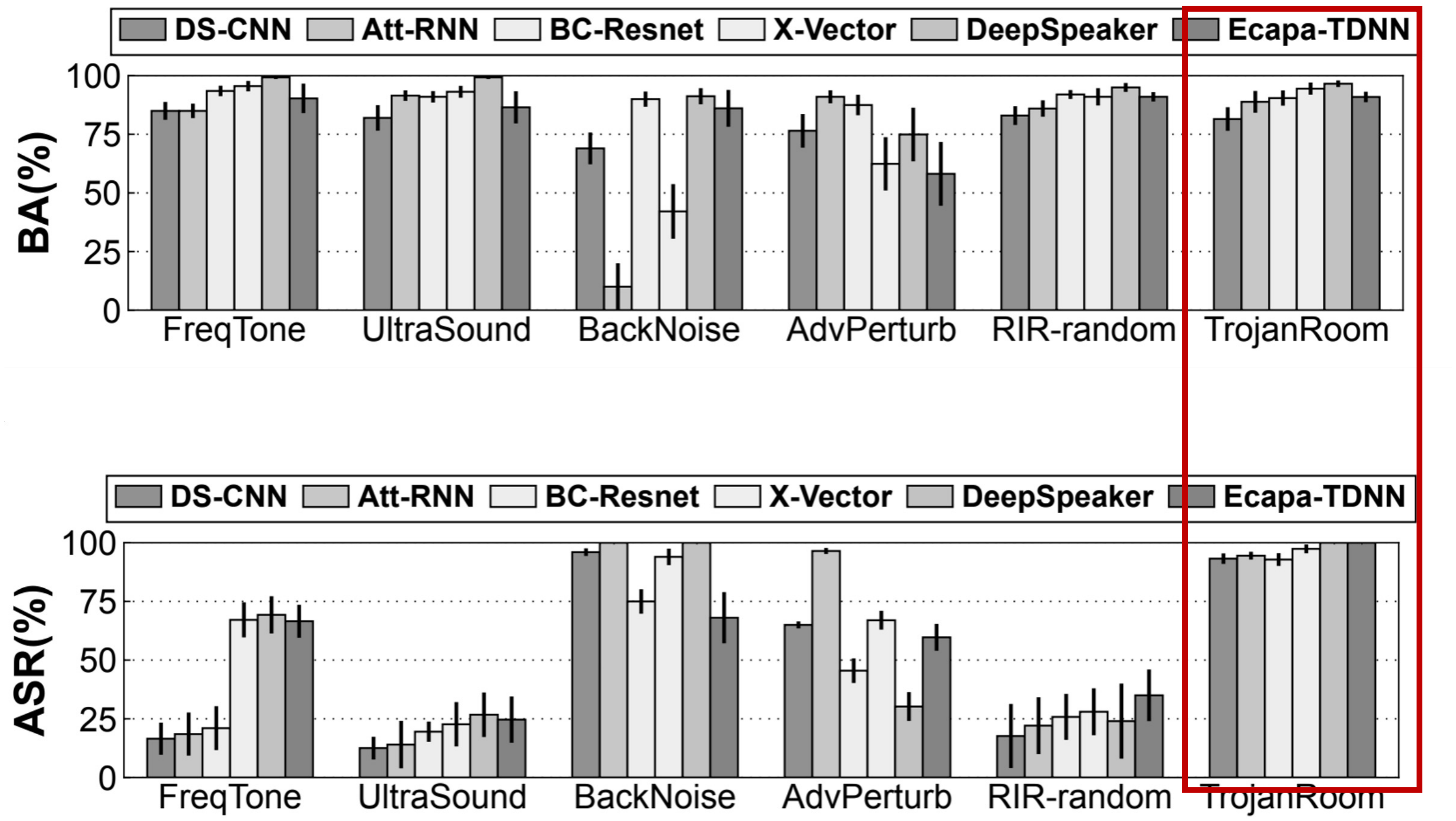
Insufficient specificity



Insufficient efficacy

Evaluation of attack efficacy and specificity

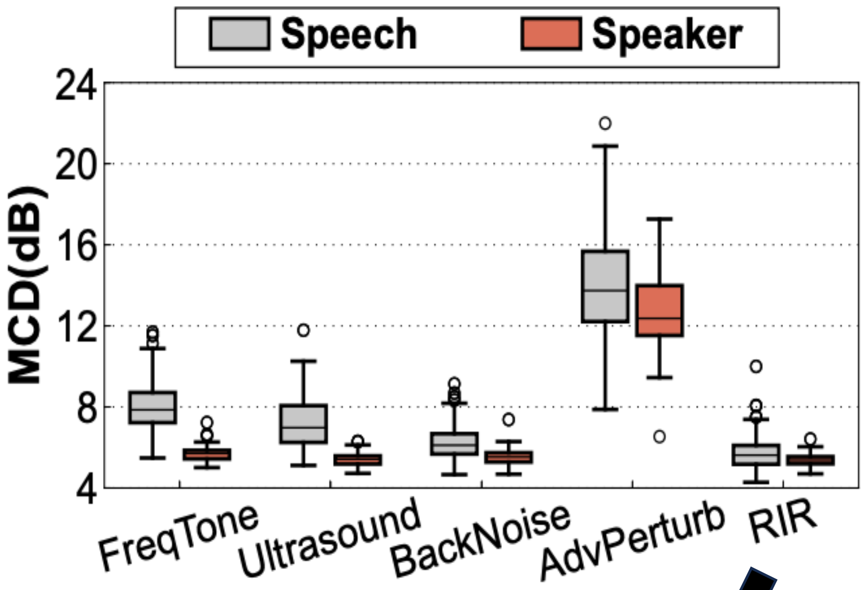
- Setup: 3 SCR models, 3 SR models, 5 baselines

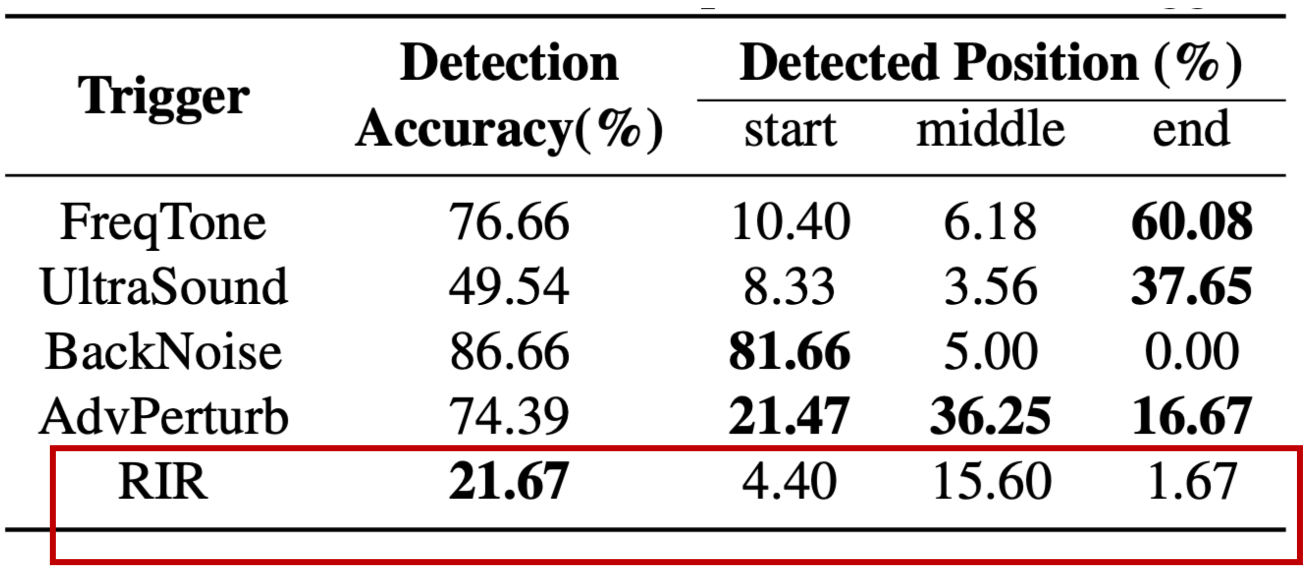


BA drop < 3%

ASR > 90%

Evaluation of attack fidelity



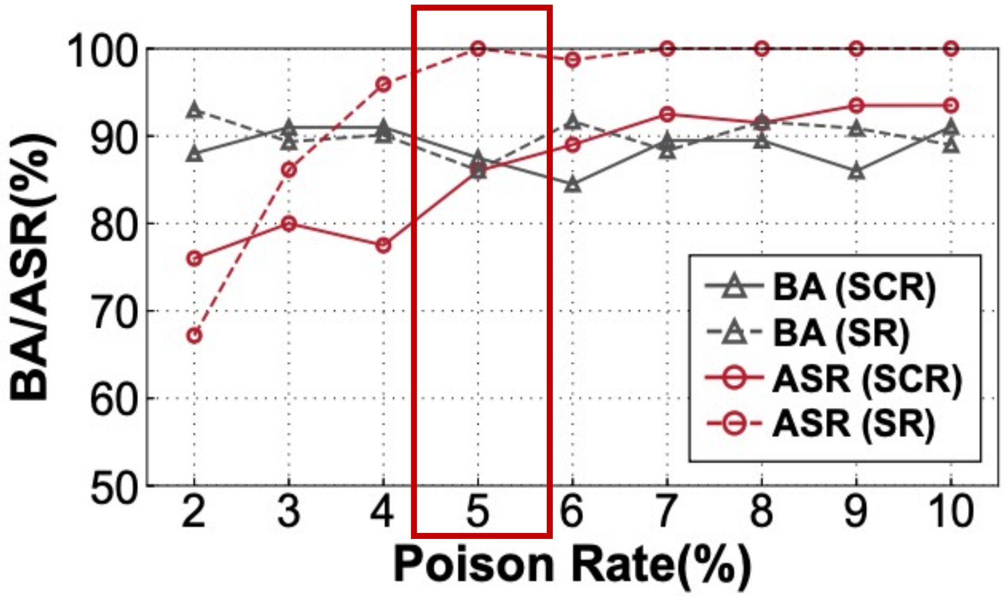


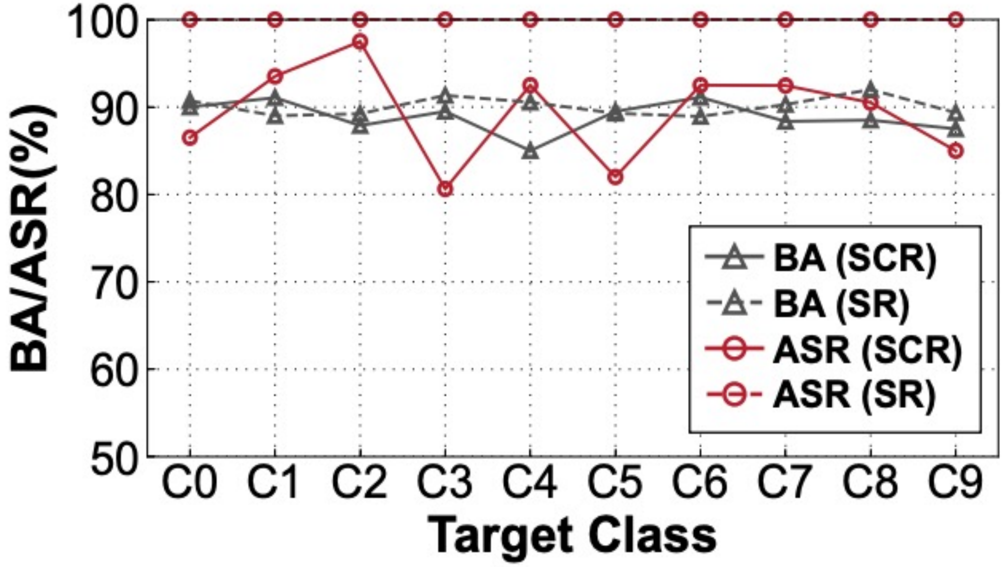
Trigger	Detection Accuracy(%)	Detected Position (%)		
		start	middle	end
FreqTone	76.66	10.40	6.18	**60.08**
UltraSound	49.54	8.33	3.56	**37.65**
BackNoise	86.66	**81.66**	5.00	0.00
AdvPerturb	74.39	**21.47**	**36.25**	**16.67**
RIR	**21.67**	4.40	15.60	1.67

RIR trigger induces less distortion between clean and poisoned speeches

Almost 80% of human listeners can not detect RIR triggers from clean speeches

Investigation of various impact factors

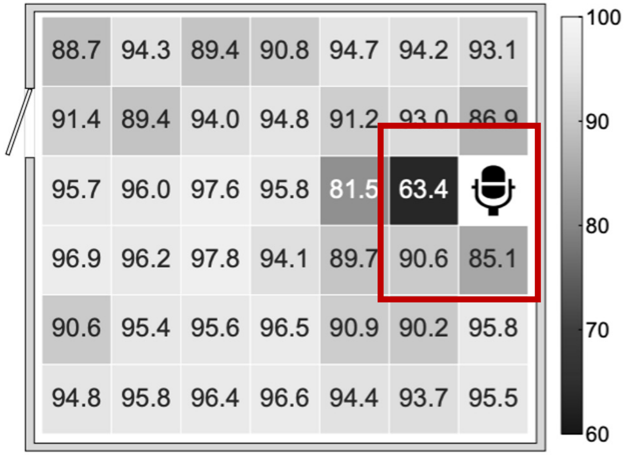




Poison rate can be reduced to 5%

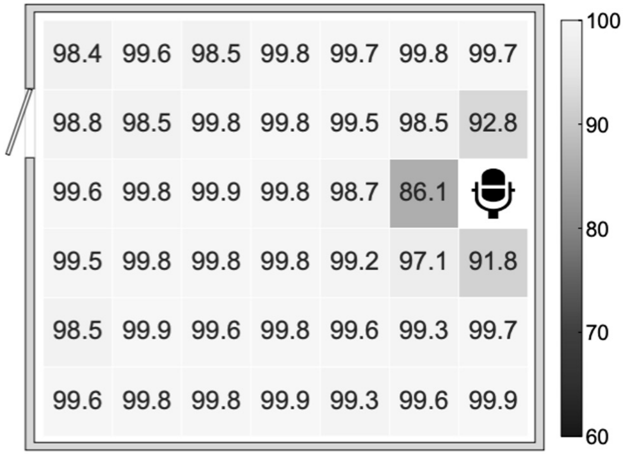
Negligible impact of different targets

Investigation of various impact factors



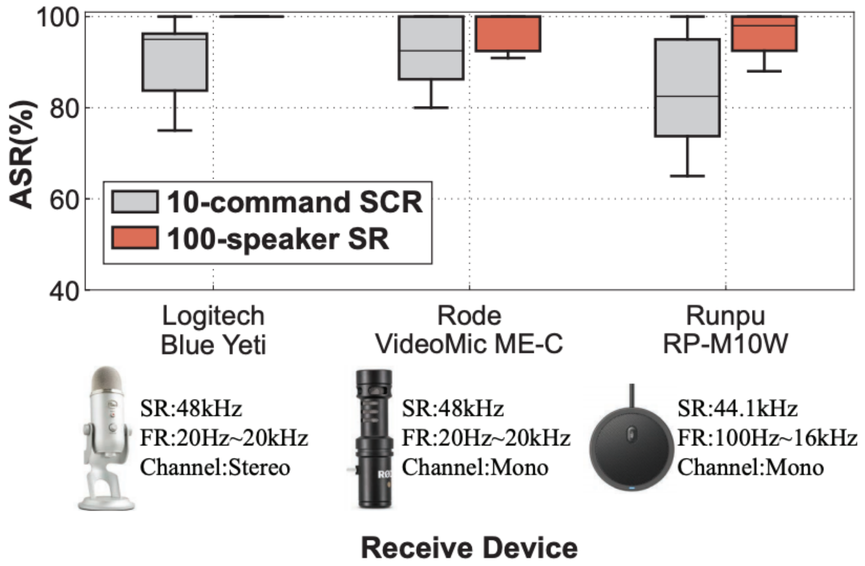
88.7	94.3	89.4	90.8	94.7	94.2	93.1
91.4	89.4	94.0	94.8	91.2	93.0	86.9
95.7	96.0	97.6	95.8	81.5	63.4	
96.9	96.2	97.8	94.1	89.7	90.6	85.1
90.6	95.4	95.6	96.5	90.9	90.2	95.8
94.8	95.8	96.4	96.6	94.4	93.7	95.5

(b) ASR of 10-command SCR



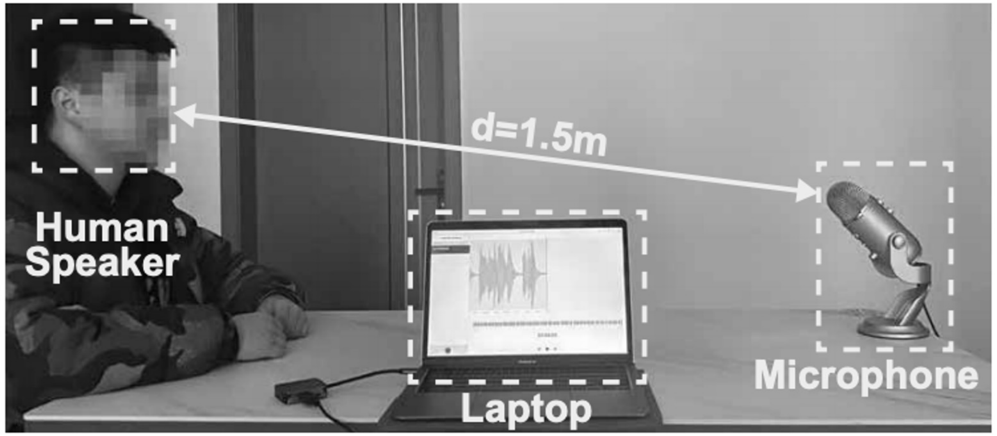
98.4	99.6	98.5	99.8	99.7	99.8	99.7
98.8	98.5	99.8	99.8	99.5	98.5	92.8
99.6	99.8	99.9	99.8	98.7	86.1	
99.5	99.8	99.8	99.8	99.2	97.1	91.8
98.5	99.9	99.6	99.8	99.6	99.3	99.7
99.6	99.8	99.8	99.9	99.3	99.6	99.9

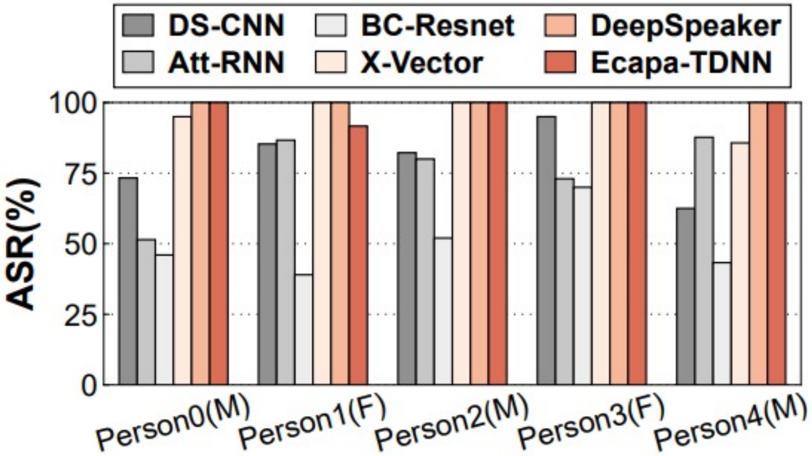
(c) ASR of 100-speaker SR



- Realize a long attack distance of 5m
- Attack degrades at a near distance due to weaker reverberation
- High-end microphones used by the audio system lead to better attack performance

Demonstration of live-speech attack





It’s practical to perform live-speech attack in real world

Countermeasures

■ **Source-level liveness detection**

- VOID and LCNN

■ **Data-level trigger disruption**

- Band-pass Filtering, Resampling, Re-quantization, and Mel Extraction-Inversion

■ **Model-level backdoor defense**

- Fine-pruning, Spectral Signature, and Neural Cleanse

Summary

- Sound channel distortion causes digital audio backdoor attacks fail
- Channel distortion itself can serve as a physical trigger
- We design a systematic method to launch the physical audio backdoor attack

Thank You