FraudWhistler: A Resilient, Robust and Plug-and-play Adversarial Example Detection Method for Speaker Recognition

Kun Wang[1], Xiangyu Xu[2], Li Lu[1], Zhongjie Ba[1], Feng Lin[1], Kui Ren[1]

[1]Zhejiang University, [2]Southeast University

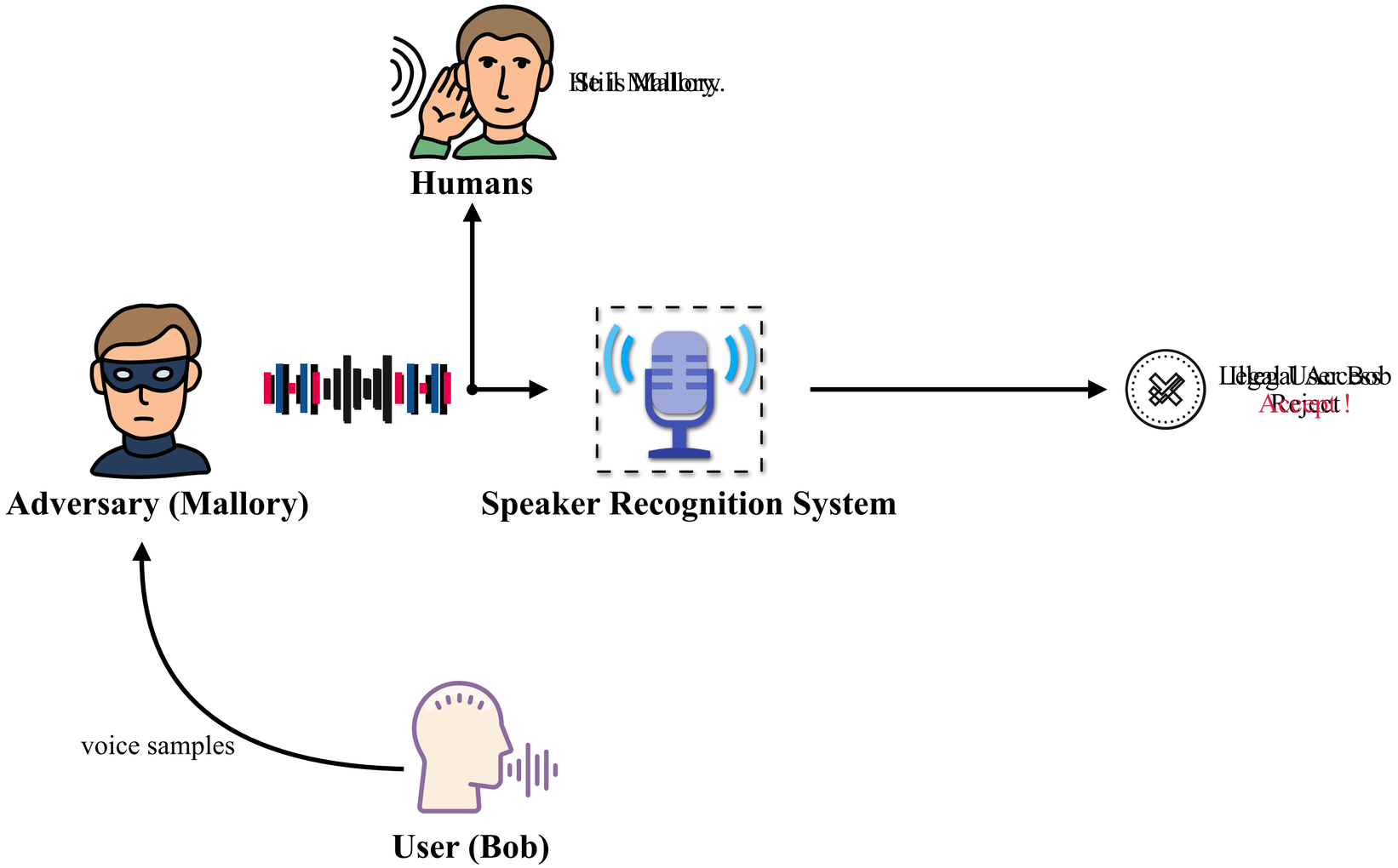
August 16, 2024

Background: Speaker Recognition (SR) has been applied widely

Definition

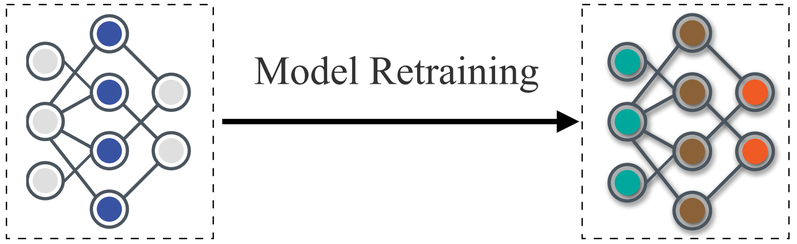
Application

Problem: DNN-based SR systems can be fooled by Adversarial Examples



Solution: Existing defenses against adversarial example attacks

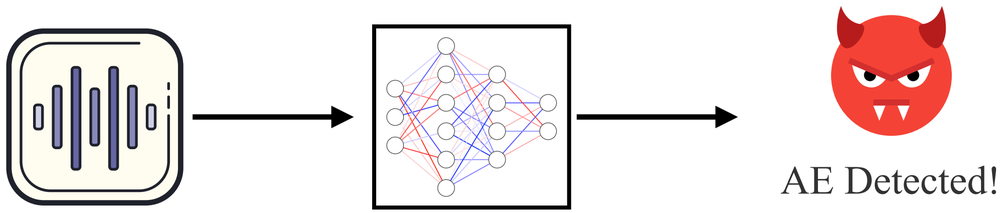
Adversarial Training



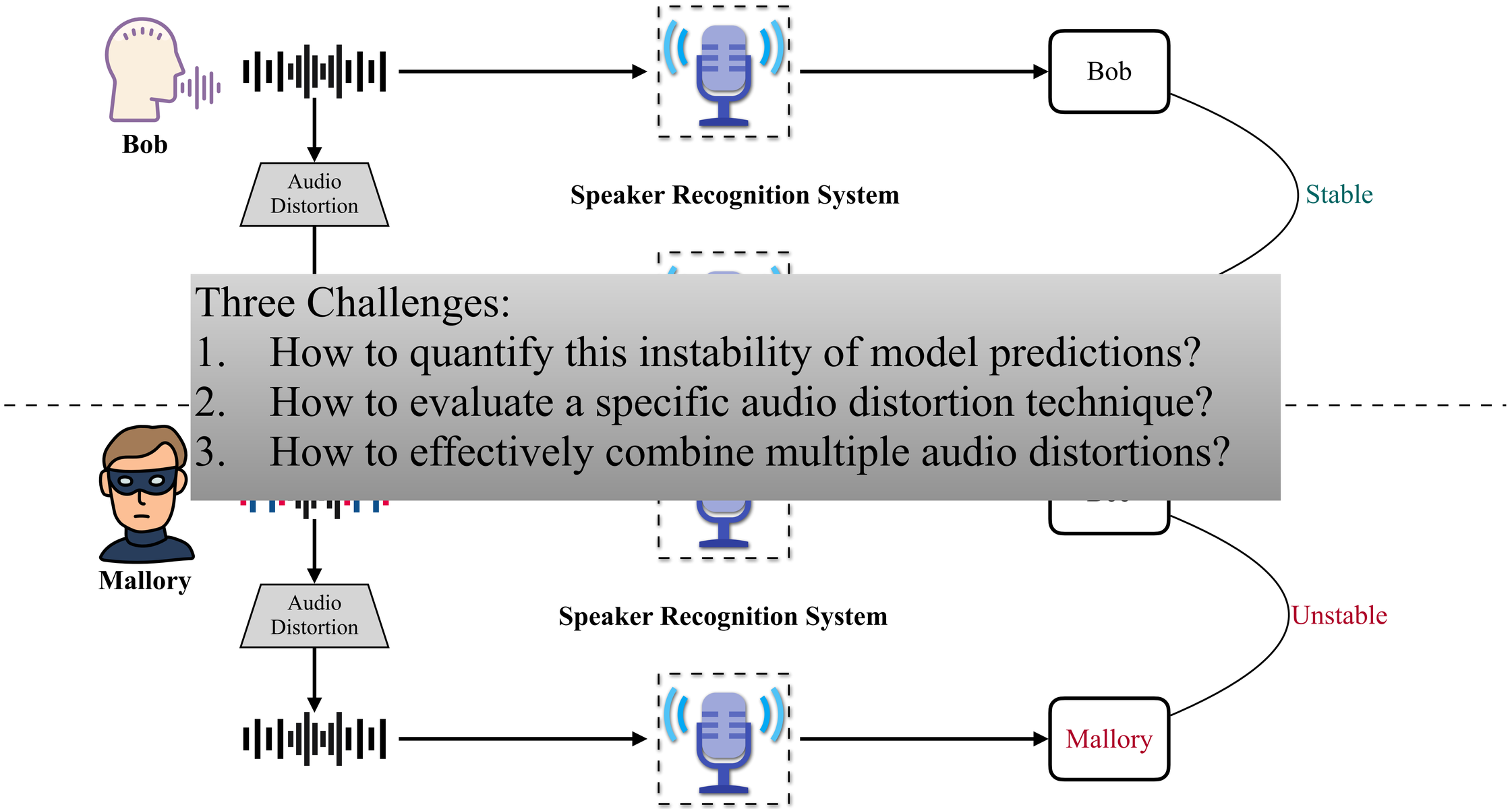
To address this issue, we proposed an **adversarial example detection method** that leverages an **intrinsic characteristic** of AEs, making it **independent of specific attack algorithms**.

Audio AE Purification

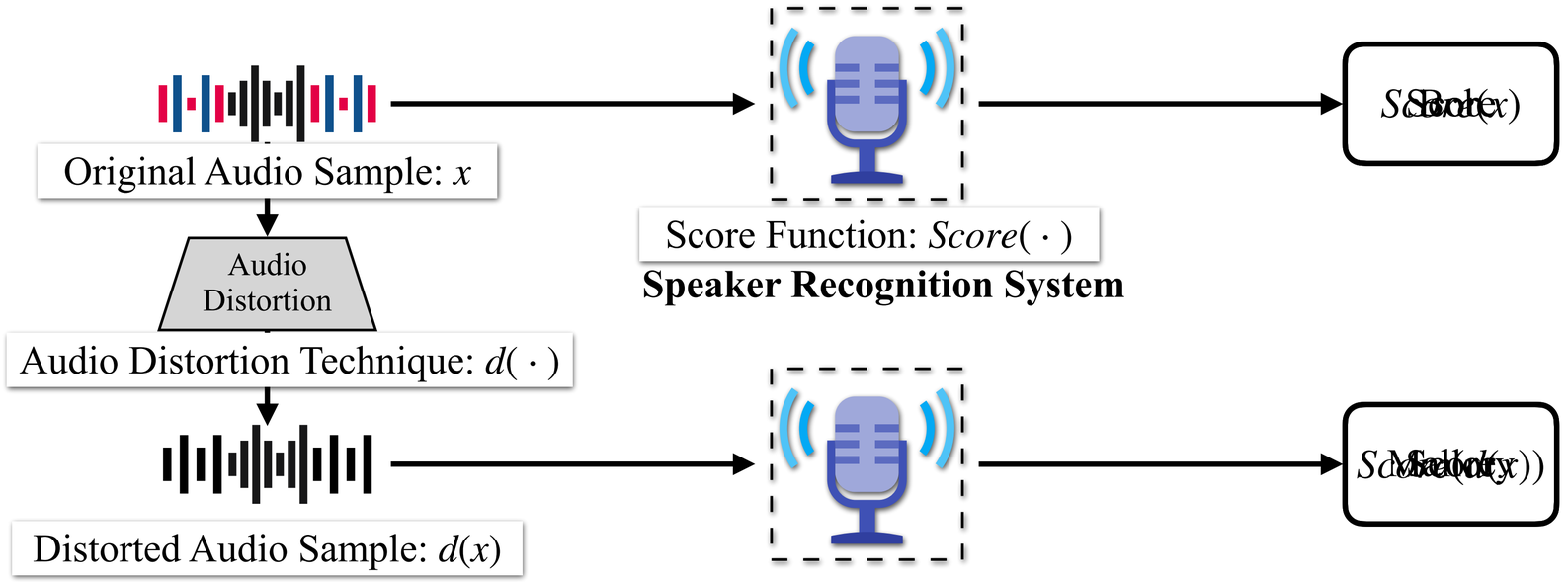
Audio AE Detection



Preliminary: An intrinsic characteristic of AEs

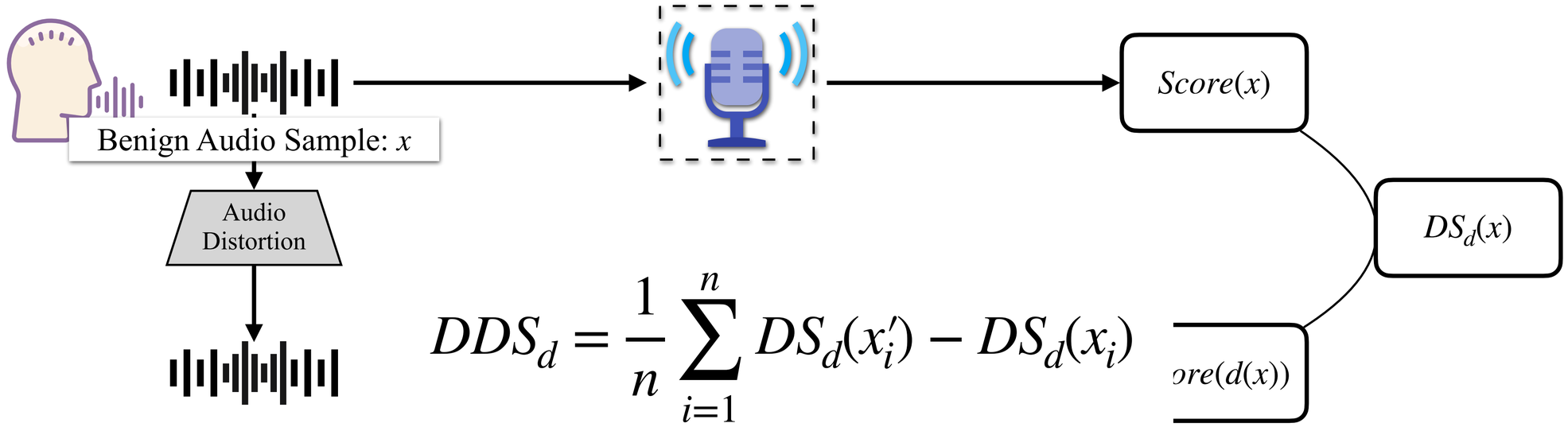


Challenge 1: How to quantify this instability of model predictions?

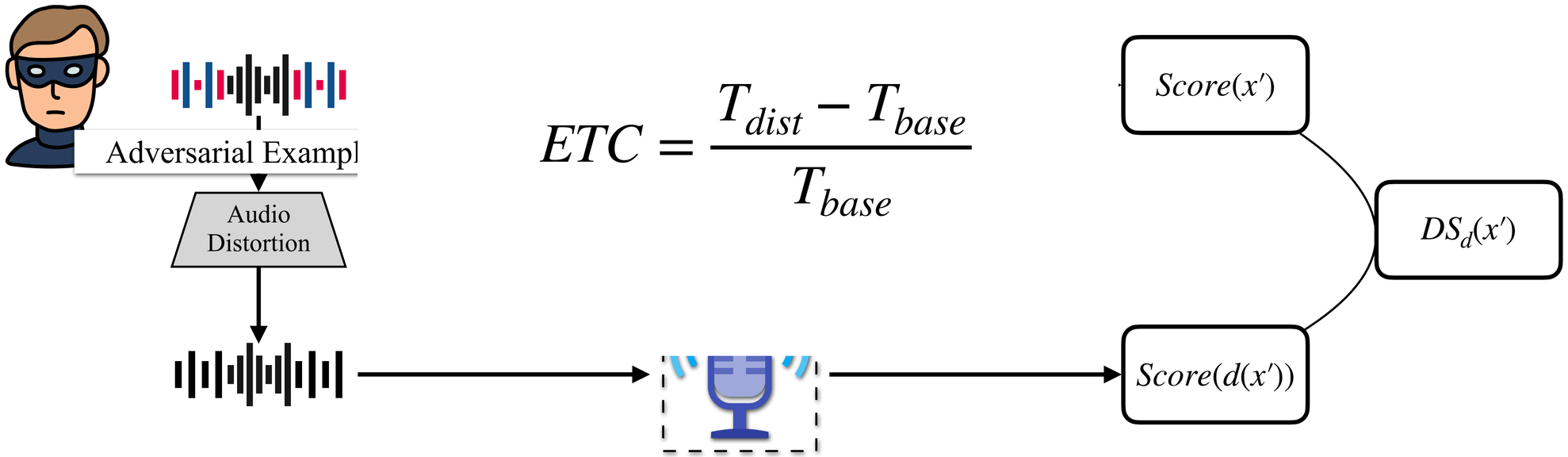


$$DS_d(x) = |Score(x) - Score(d(x))|$$

Challenge 2: How to evaluate a specific audio distortion technique?



$$DDS_d = \frac{1}{n}\sum_{i=1}^{n} DS_d(x_i') - DS_d(x_i)$$

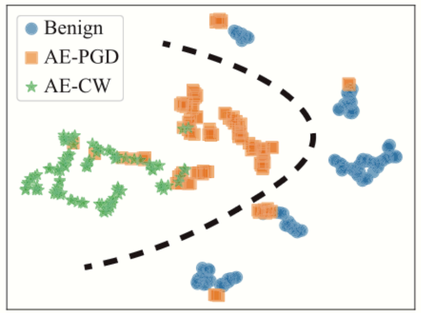


$$ETC = \frac{T_{dist} - T_{base}}{T_{base}}$$

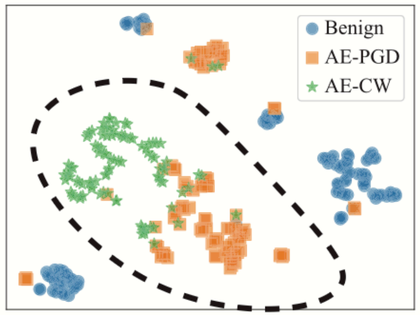
Challenge 2: How to evaluate a specific audio distortion technique?

Type	Distortion	Description	DDS_d	DDS'_d	ETC
Signal Processing	Quantization	Quantize each data point and then convert back by De-Quantization	0.61	0.24	$\approx$ .001
	Codec	Compress audio sample and decompress	0.64	0.23	$\approx$ .026
	Resample	Downsample the audio wave and upsample to original sample rate	0.34	-0.09	$\approx$ .036
	Filtering	Filter the audio wave with high-pass and low-pass filters	0.18	-0.16	$\approx$ .031
	Smoothing	Smooth the audio wave with specific window length	0.43	-0.20	$\approx$ .007
Audio Augmentation	Noisifier	Add white noise with given SNR	0.68	0.59	$\approx$ .006
	Reverber	Add reverberation effect with given RIR	0.47	0.27	$\approx$ .113
	TimeScale	Scale the speed of audio	0.34	-0.18	$\approx$ .039
	Clip	Clip the audio wave amplitude to certain range	0.07	-0.12	$\approx$ .001
	DropChunk	Drop some chunks from audio wave	0.38	0.15	$\approx$ .031
	DropFreq	Drop some frequency components from audio wave	0.39	0.18	$\approx$ .224
	PitchShift	Shift the pitch level of the audio sample	0.42	-0.06	$\approx$ 36.3
	TimeShift	Shift specific ratio of audio wave	0.11	-0.001	$\approx$ .001
Audio Reconstruction	GriffinLim	Extract MelSpectrogram and reconstruct with GriffinLim [18]	0.40	0.48	$\approx$ 295
	LPC	Extract LPC coefficients and reconstruct with random excitation	0.17	0.16	$\approx$ 16.1

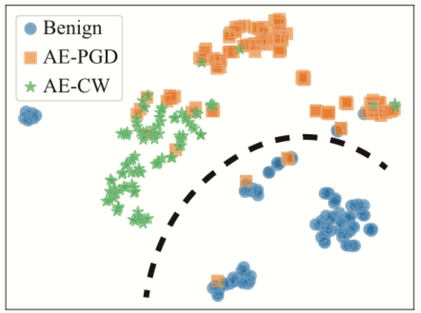
Challenge 3: How to effectively combine multiple audio distortions?



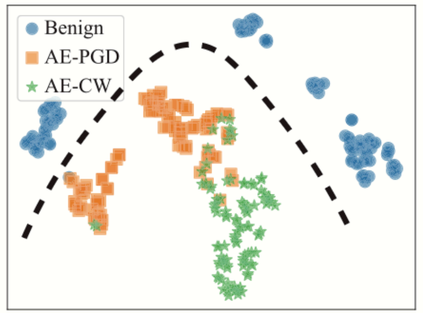
(a) 2 distortions



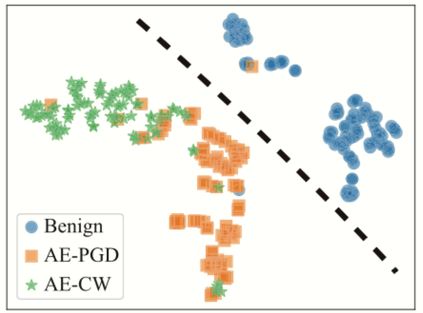
(b) 3 distortions



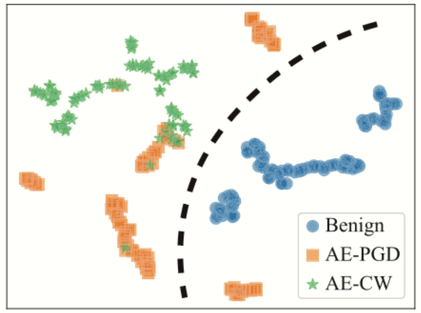
(c) 4 distortions



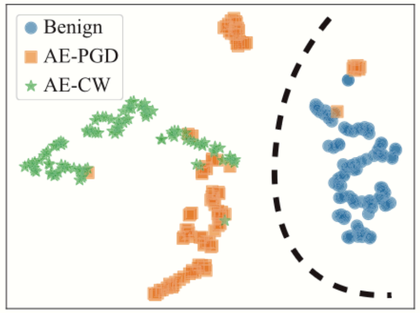
(d) 5 distortions



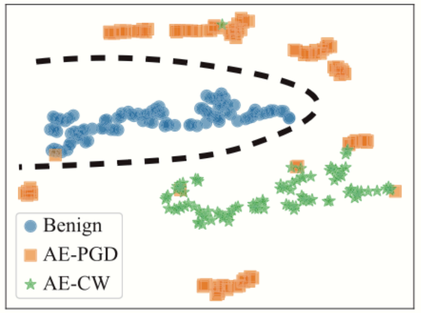
(e) 6 distortions



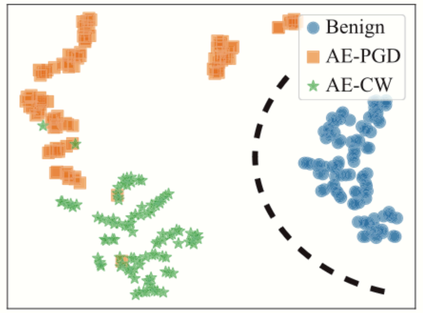
(f) 2 distortion levels



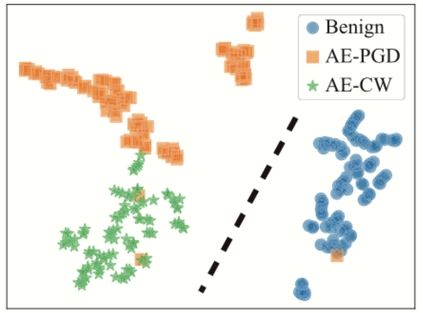
(g) 3 distortion levels



(h) 4 distortion levels

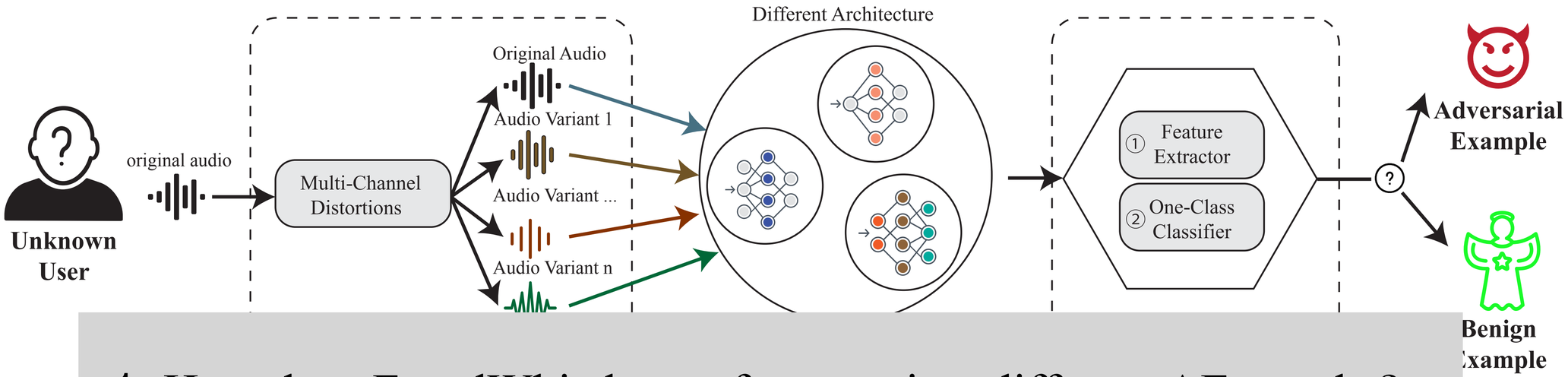


(i) 5 distortion levels



(j) 6 distortion levels

Method: Framework of FraudWhistler



1. How does FraudWhistler perform against different AE attacks?
2. How does FraudWhistler perform in physical world?
3. How does FraudWhistler perform against an adaptive adversary?

Result 1: How does FraudWhistler perform against different AE algorithms?

Experimental Setup

- Dataset: VCTK (**110** speakers)
- SR system: ECAPA-TDNN[1] (VoxCeleb 1/2)
- **5 AE Algorithms**: FGSM, PGD, CW, FM, UNIV
- SOTA Detections: WG[2], TD[3]
- Running Environment:
  - Ubuntu hirsute 21.04
  - 40 Intel Xeon Silver 4210R CPU
  - 256 RAM
  - Four 48GB NVIDIA RTX A6000 GPU

- Evaluation Metrics
  - Detect Accuracy on AEs (ACC_{ae}):

$$\frac{\#\text{Detected Adversarial Examples}}{\#\text{Total Adversarial Examples}}$$

  - Accuracy on Benign Examples (ACC_{be}):

$$\frac{\#\text{Accepted Benign Examples}}{\#\text{Total Benign Examples}}$$

  - Robust Accuracy on AEs (ACC_{rob}):

$$1 - \frac{\#\text{Successful Adversarial Examples}}{\#\text{Total Adversarial Examples}}$$

[1] B. Desplanques et al. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In Proc. of INTERSPEECH, 2020.
[2] S. Hussain et al. WaveGuard: Understanding and Mitigating Audio Adversarial Examples. In Proc. of USENIX Security, 2021.
[3] Z. Yang et al. Characterizing Audio Adversarial Examples Using Temporal Dependency. In Proc. of ICLR, 2018.

Result 1: How does FraudWhistler perform against different AE algorithms?

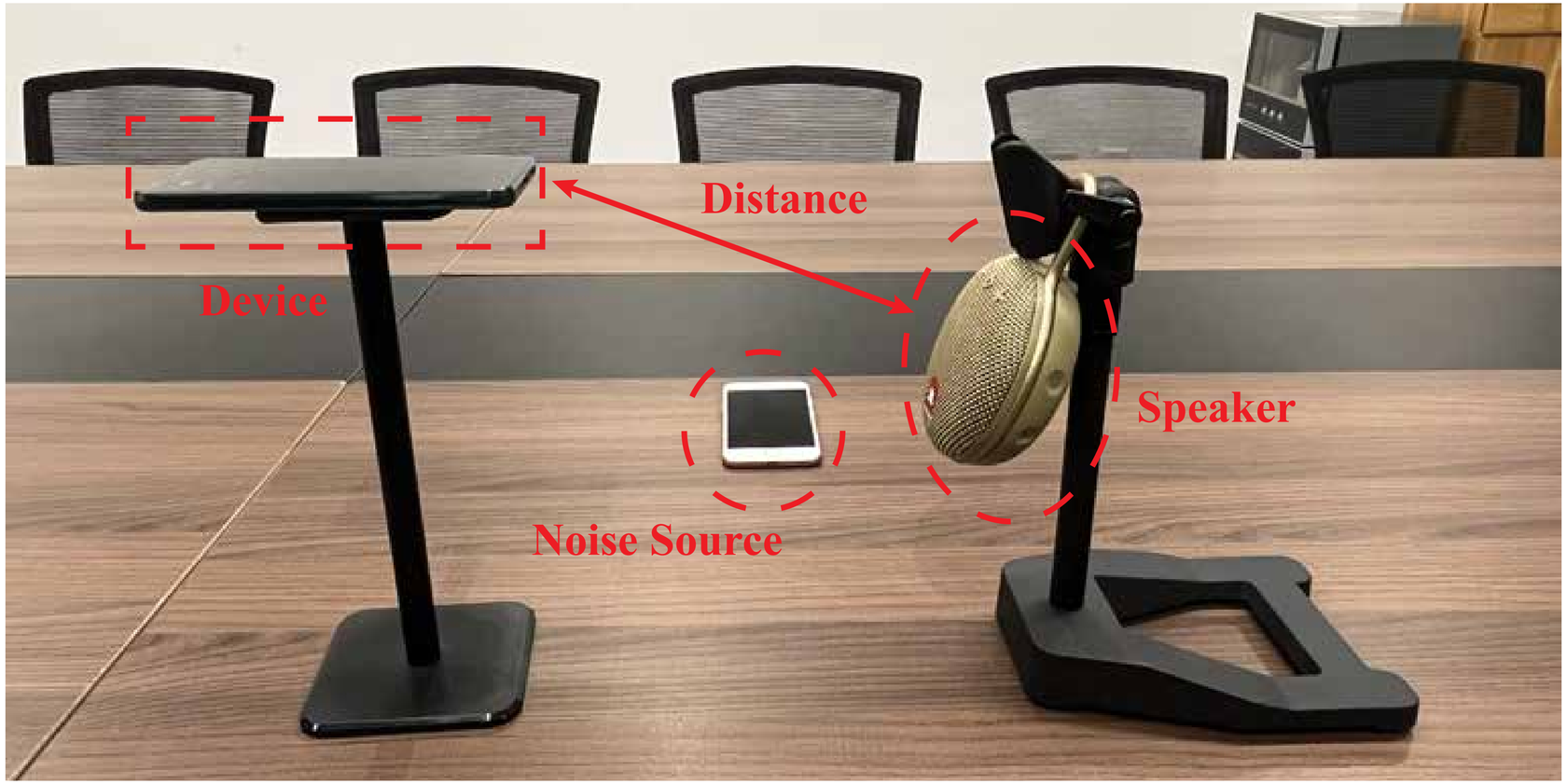
Overall Performance

SR Architectures	$ACC_{ae}(\%)$			$ACC_{be}(\%)$				$ACC_{rob}(\%)$			
	FW	WG	TD	No Defense	FW	WG	TD	No Defense	FW	WG	TD
ASV	**87.8**	80.6	43.6	99.67	94.33	92.00	**94.67**	16.6	**97.6**	80.6	58.6
CSI	**86.2**	69.0	45.6	100.0	94.33	**96.00**	92.00	23.4	**100**	90.8	64.4
OSI	**94.0**	86.0	49.6	100.0	92.67	**94.67**	93.33	17.2	**98.4**	86.0	61.4

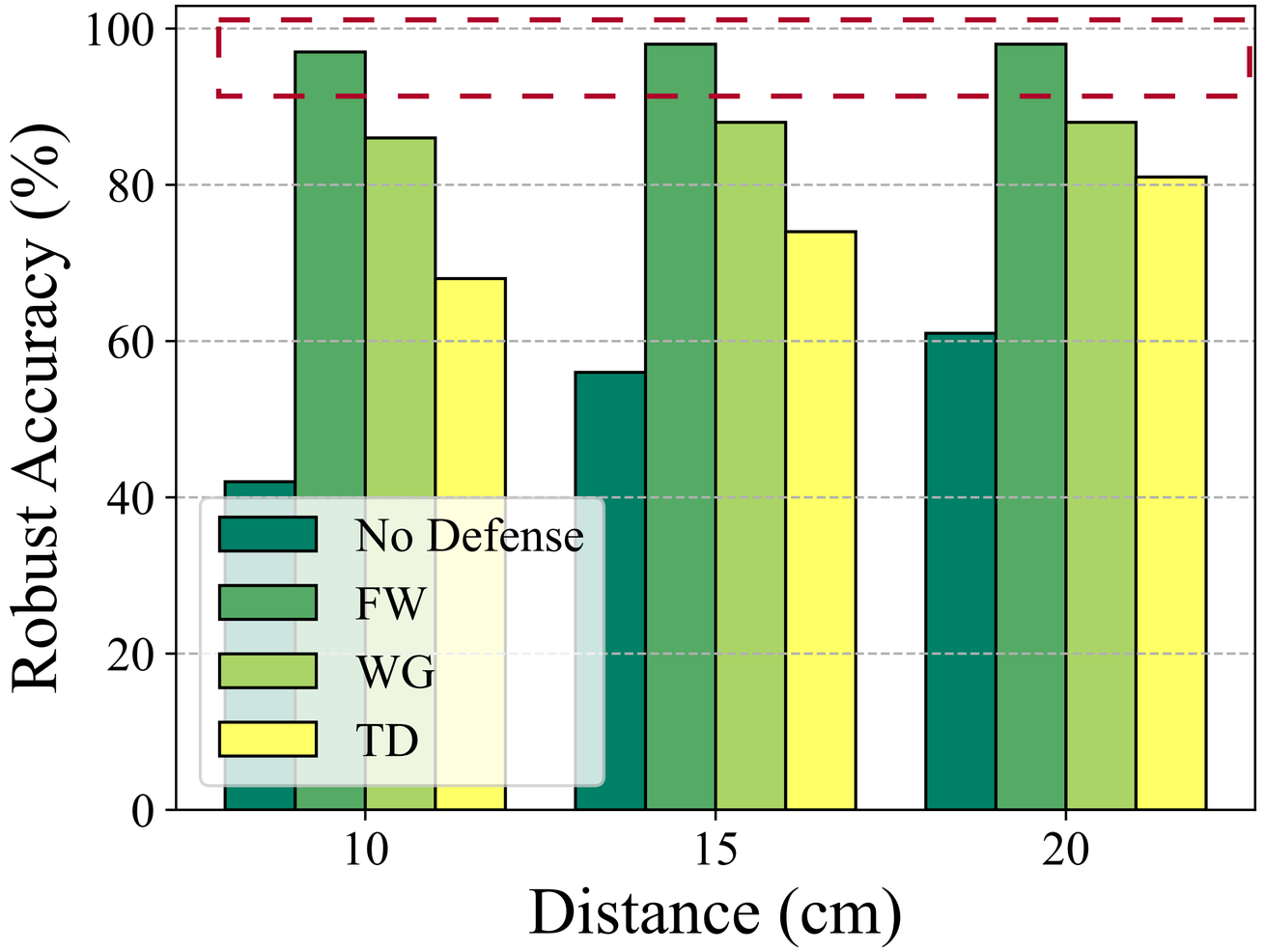
- FraudWhistler achieves highest detection accuracy on AEs for all SR architectures.
- FraudWhistler only induces a minor degradation of **6.1%** accuracy on benign examples.
- The SR system protected by FraudWhistler achieves **97.6%** robust accuracy at worst compared to **80.6%** for WG and **58.6%** for TD.

Result 2: How does FraudWhistler perform in physical world?

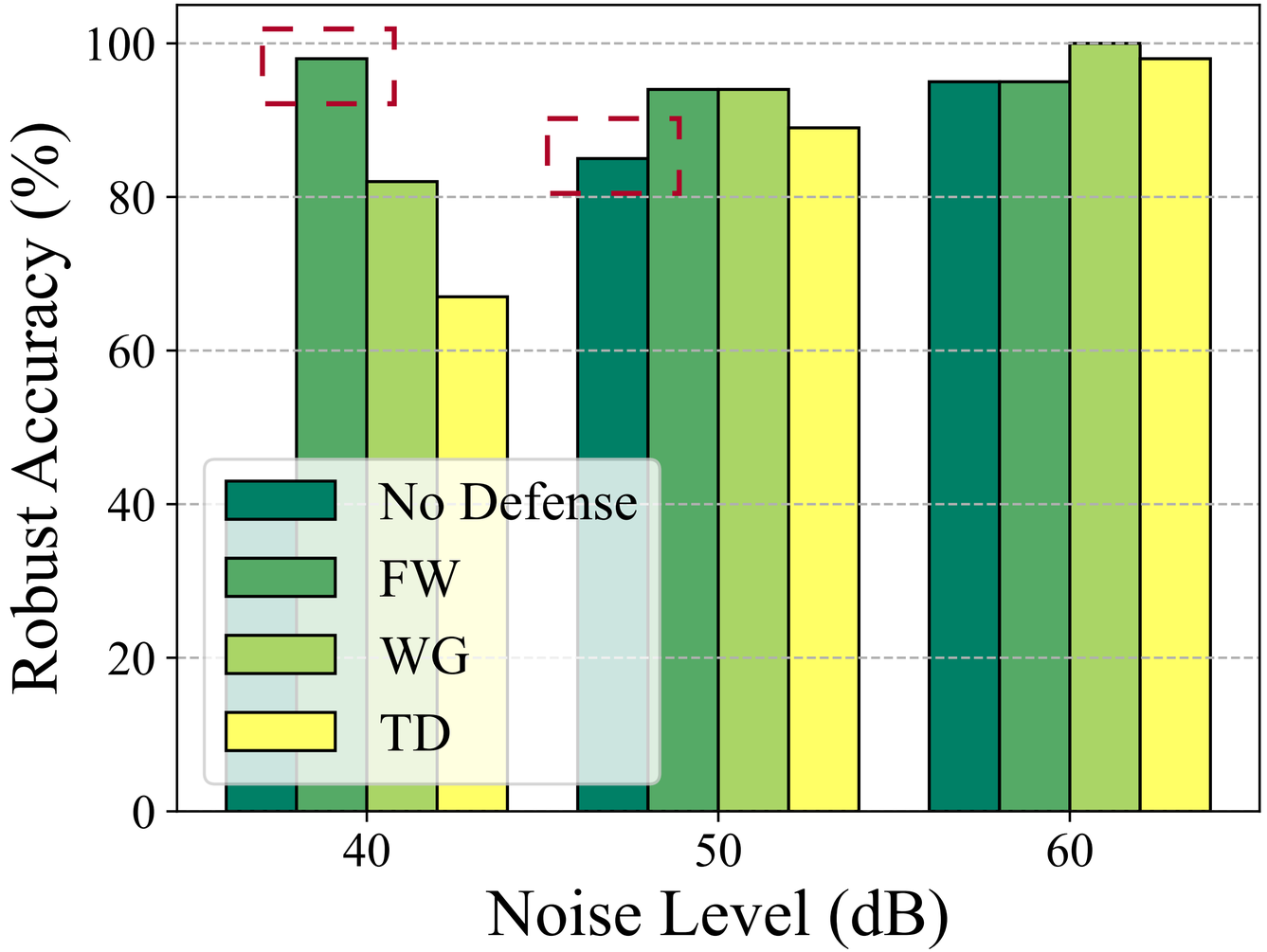
Experimental Setup



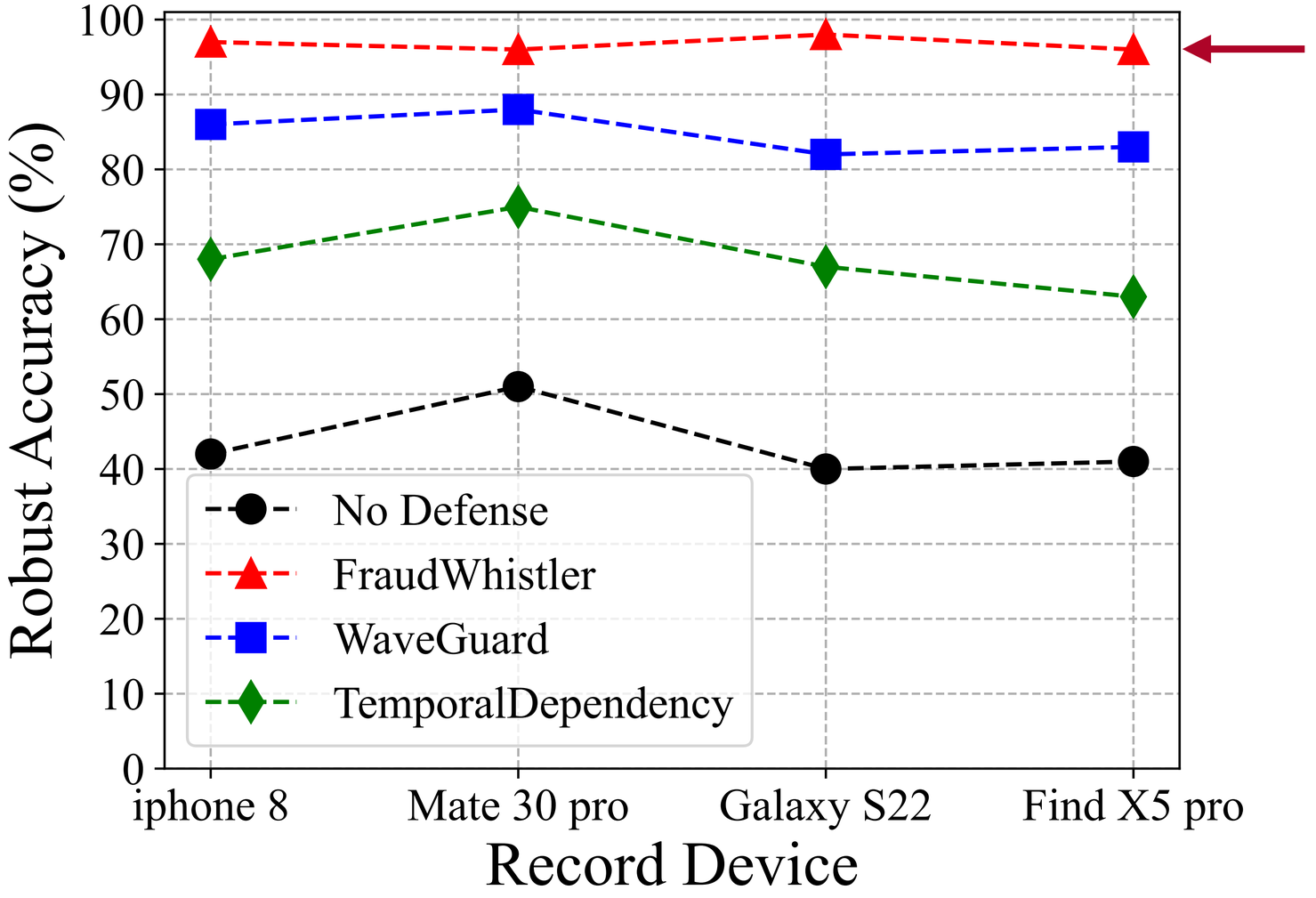
Result 2: How does FraudWhistler perform in physical world?



Result 2: How does FraudWhistler perform in physical world?



Result 2: How does FraudWhistler perform in physical world?

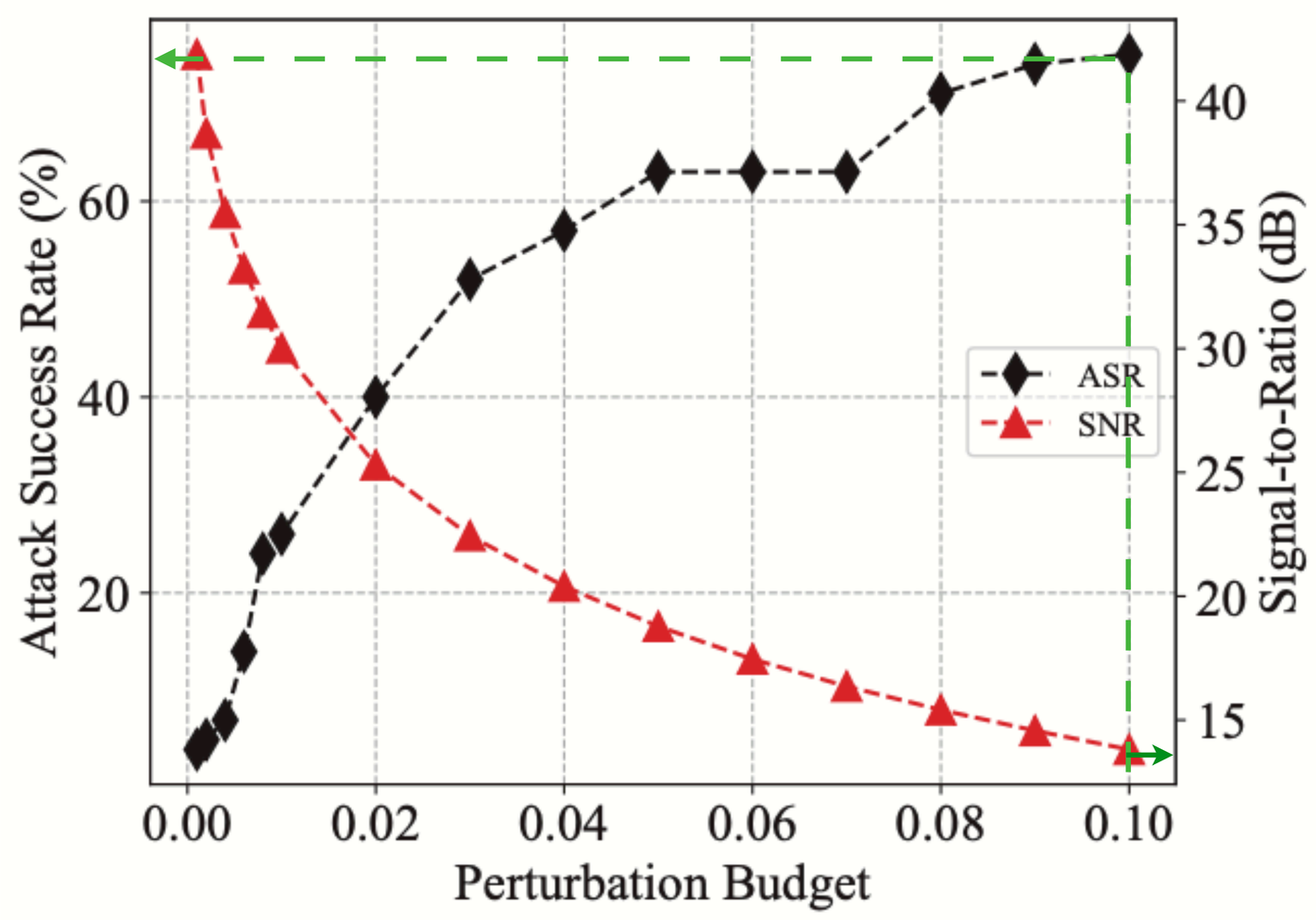


Result 3: How does FraudWhistler perform against an adaptive adversary?

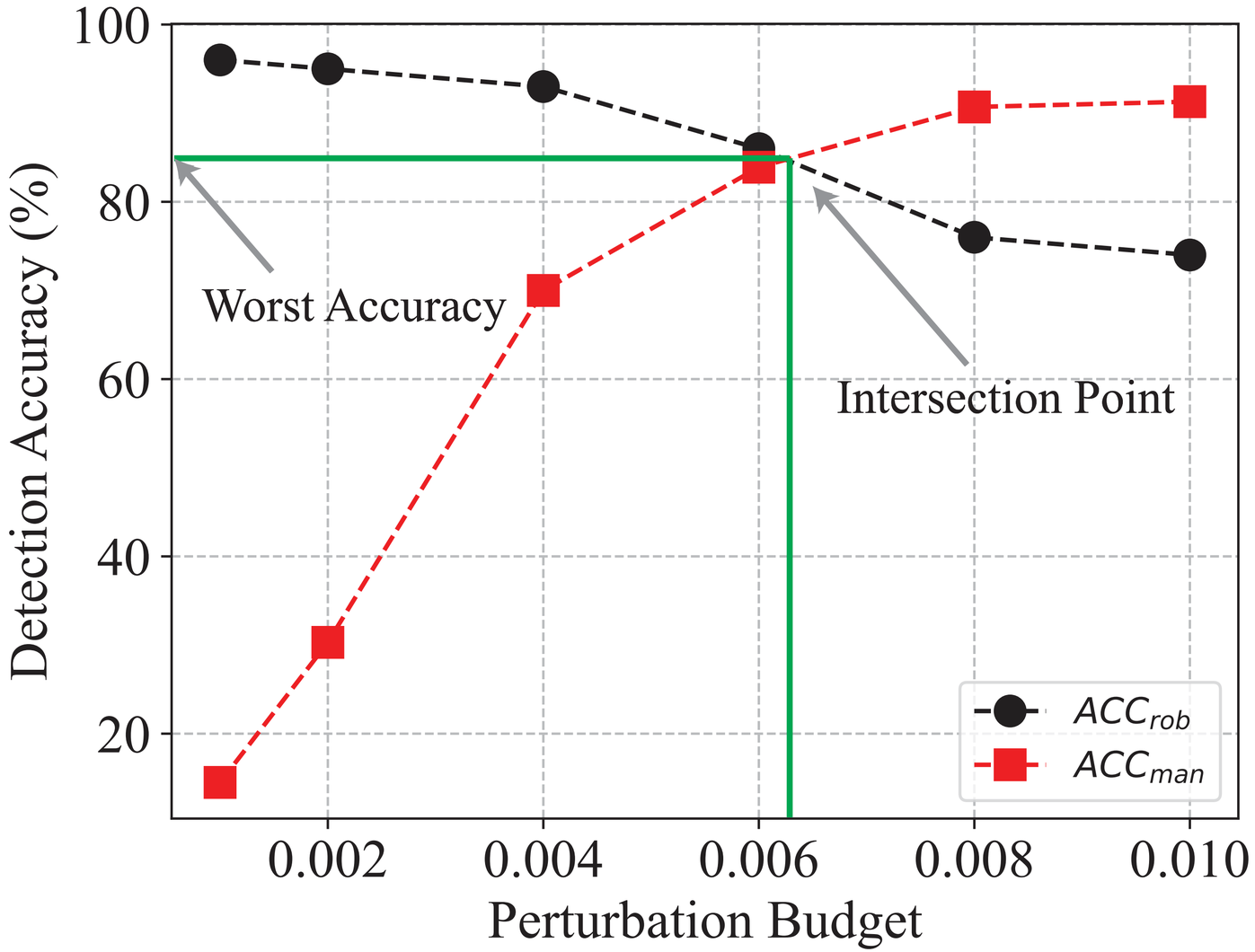
$$min\ \mathcal{L}(SR(x+\delta),t) + \sum_{i=1}^{6} c_i \cdot \mathcal{L}(SR(d_i(x+\delta)),t)$$

$$s.t.\ \|\delta\|_\infty < \varepsilon,$$

Result 3: How does FraudWhistler perform against an adaptive adversary?



Result 3: How does FraudWhistler perform against an adaptive adversary?



Summary

- We present a study evaluating a wide range of audio distortion techniques from several aspects.
- We propose an audio distortion-based adversarial example detection method for speaker recognition systems, FraudWhistler.
- The experimental results show that FraudWhistler achieves resilient performance against various attack algorithms, robust in complex realistic conditions, and effective even under adaptive attack settings.

Thanks !