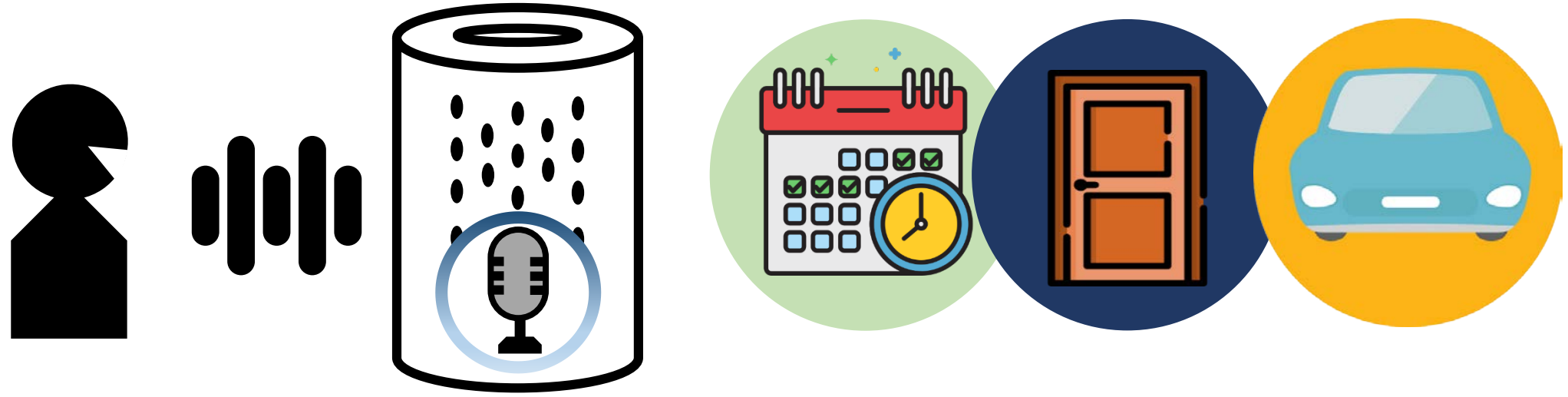
Detecting Audio Adversarial Examples with Logit Noising

Namgyu Park Sangwoo Ji Jong Kim

High Performance Computing Lab (**HPC**)
Pohang University of Science and Technology, South Korea
POSTECH

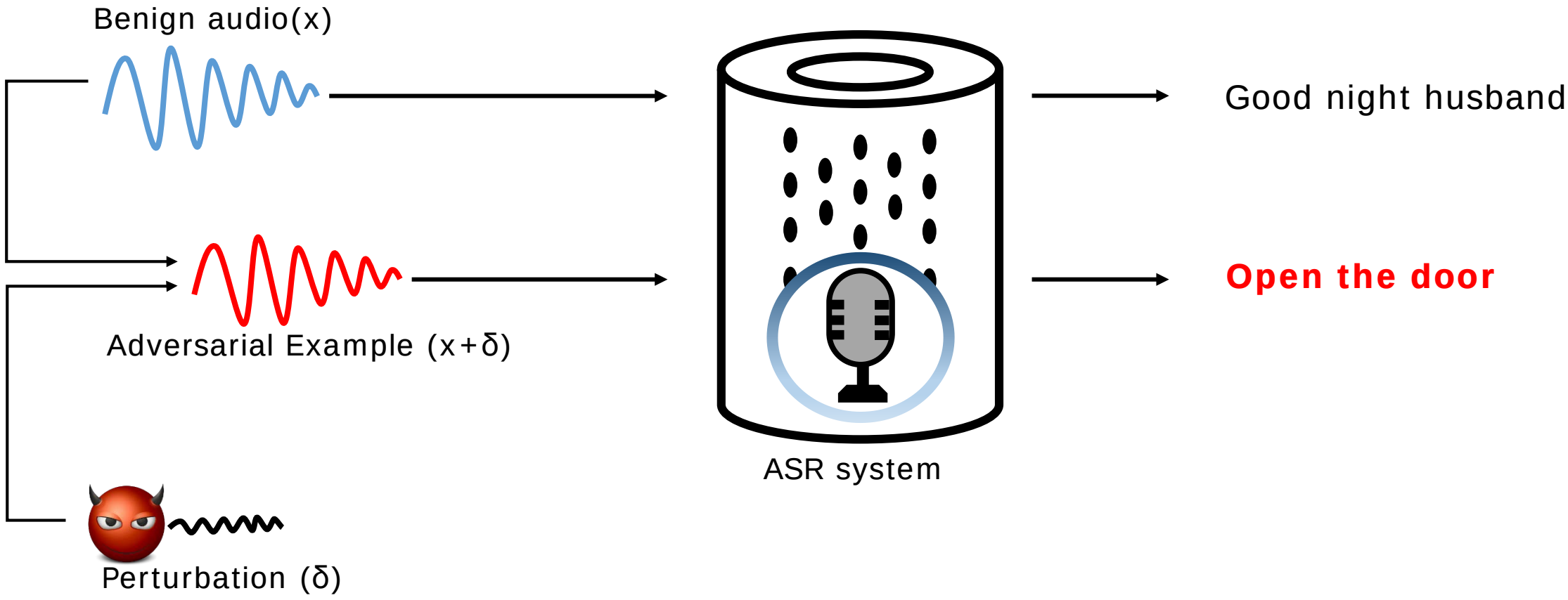
The Ubiquitous Automatic Speech Recognition (ASR)

- ASR systems have been widely used in various device



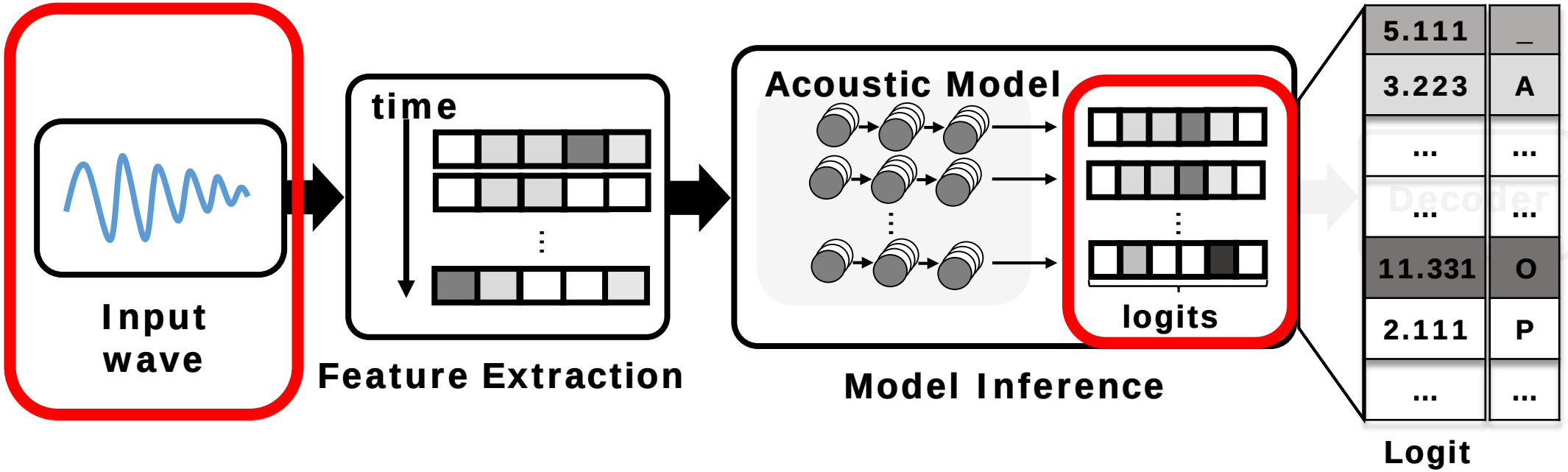
Audio Adversarial Examples

- Slightly **modified input** that fools an ASR system



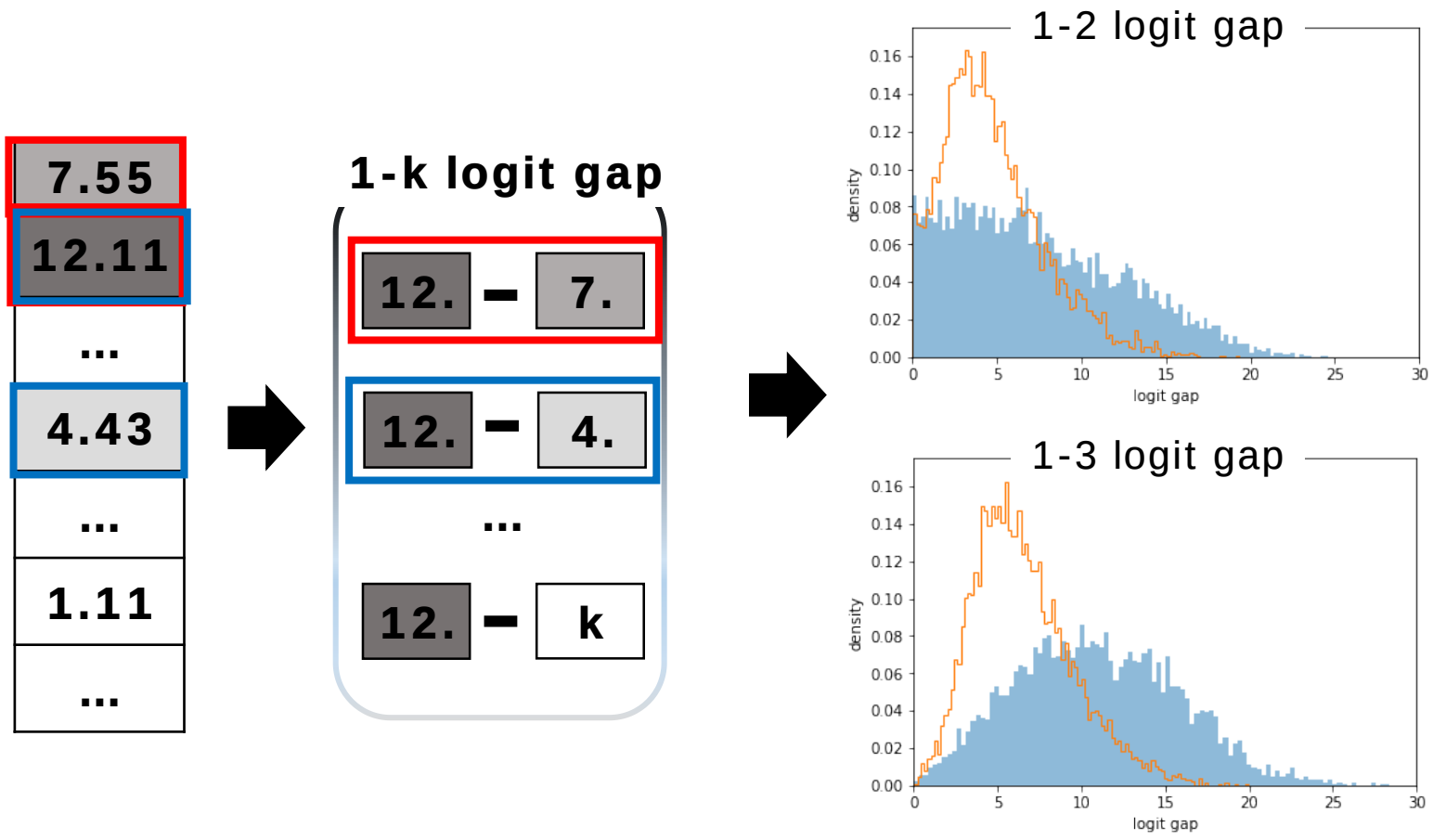
Defenses for ASR Systems

- What is the difference between attacks and benign samples?



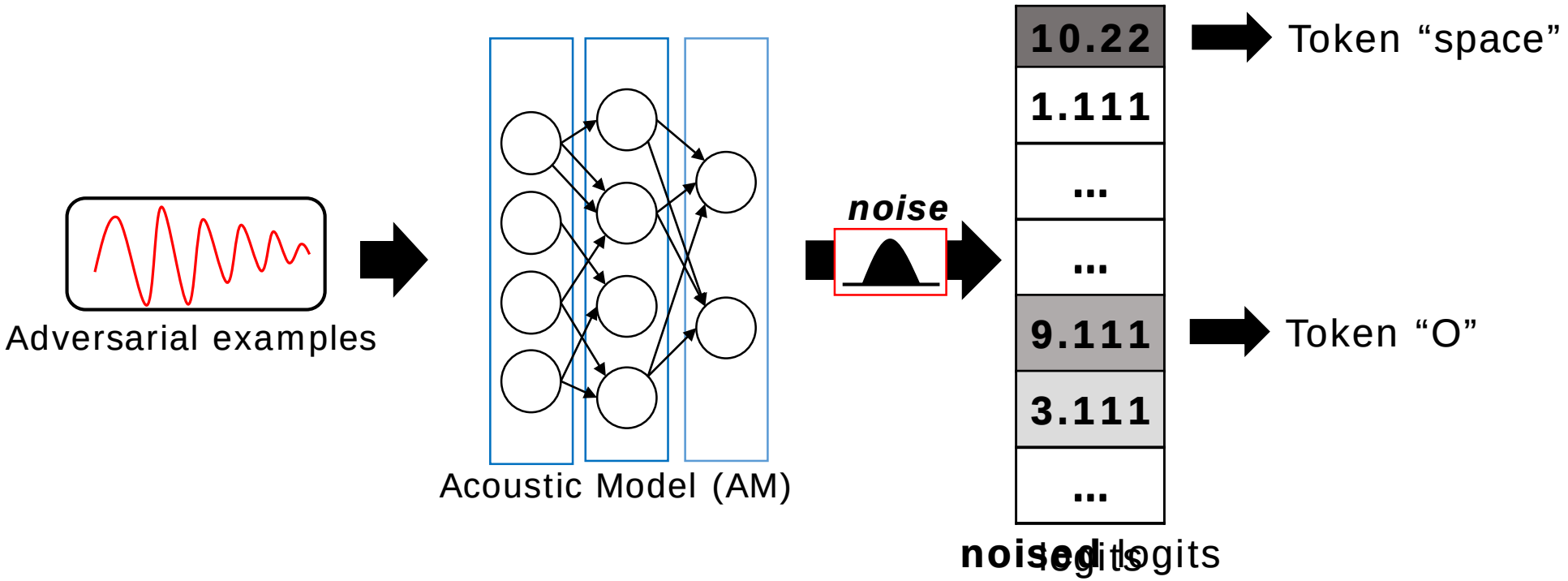
Observation: Logit Gap Distribution

- Analysis of the logit gap distribution

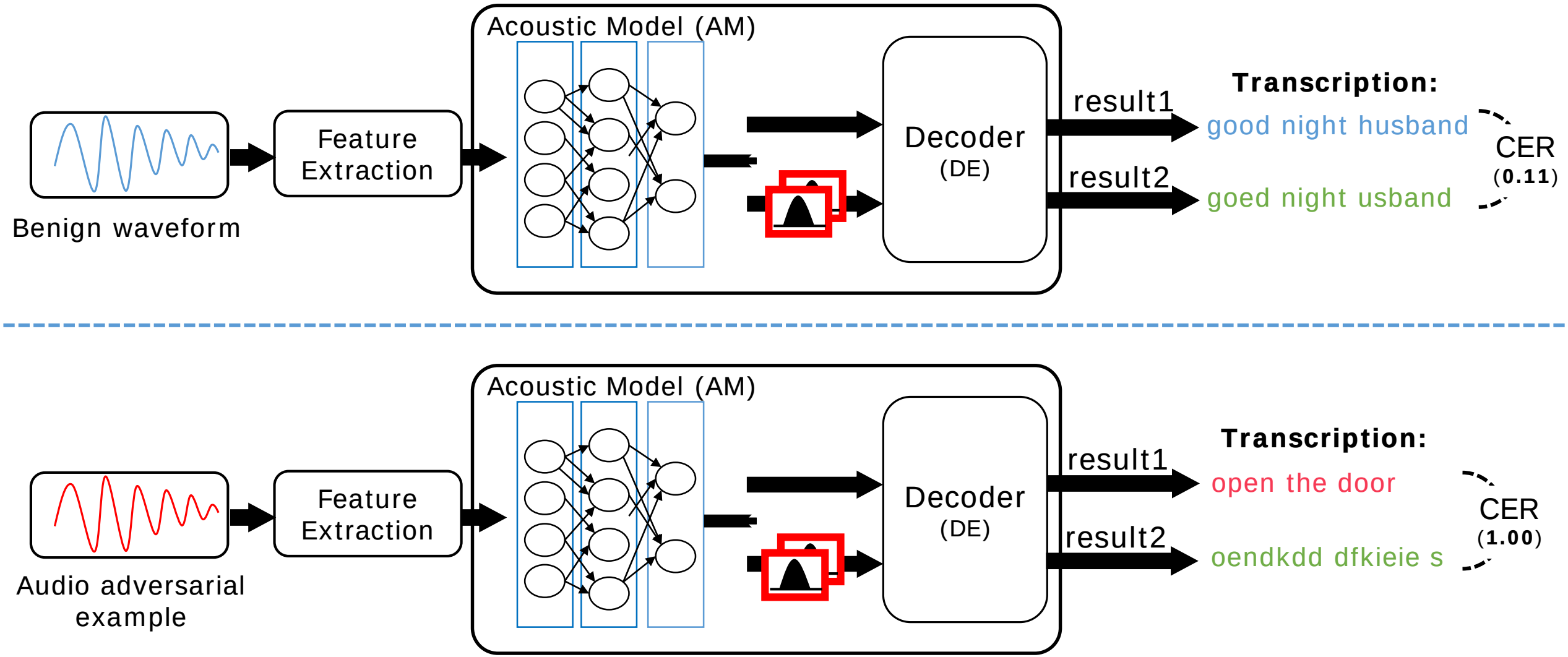


Proposed Method: Logit Noising

- When a certain random noise is added to logits, adversarial logits are easily inverted.



Logit Noising Architecture



Deciding Adversarial Examples

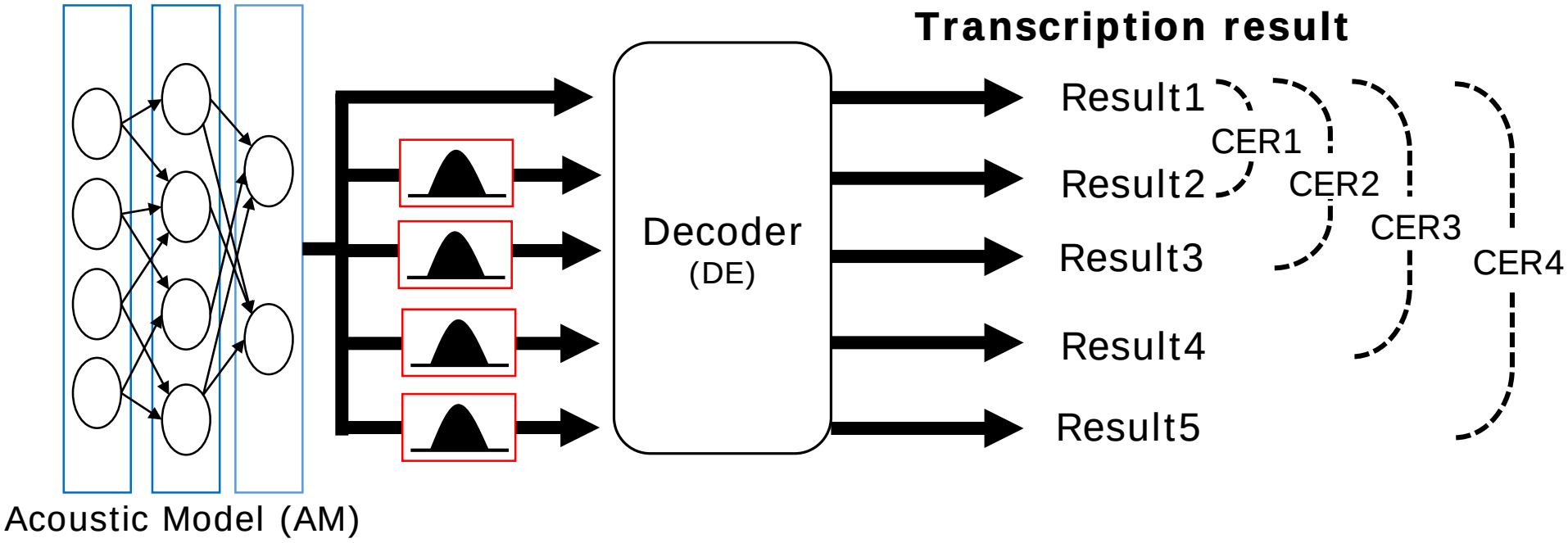
- Input audio is classified as adversarial if the character error rate (CER) differs significantly

$$\textbf{\textit{CER}}(\mathrm{DE}(\mathrm{AM}(x)), \mathrm{DE}(\mathrm{AM}(x) + \delta)) > \mathbf{t}$$

- Acoustic model(AM)
- Decoder(DE)
- Given audio(x)
- Gaussian noise(δ)

Leveraging Multiple Instances

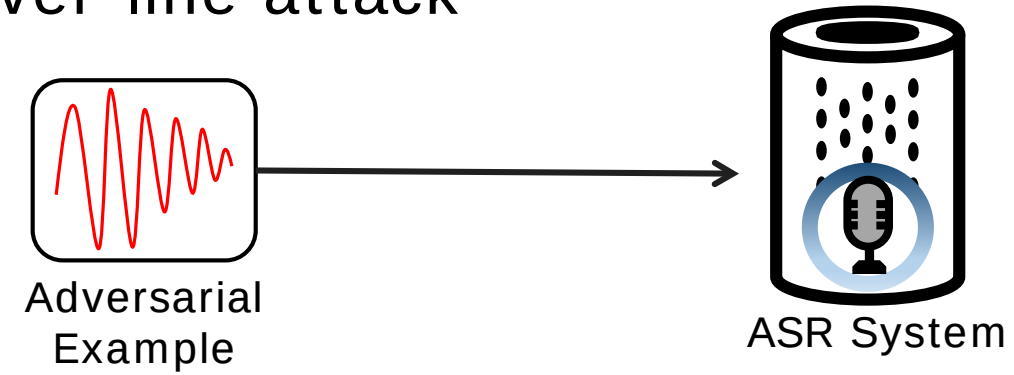
- Gaussian noise has randomness, the average result of multiple instances increases the detection accuracy



$$\frac{(CER1 + CER2 + CER3 + CER4)}{4} > \text{Threshold}$$

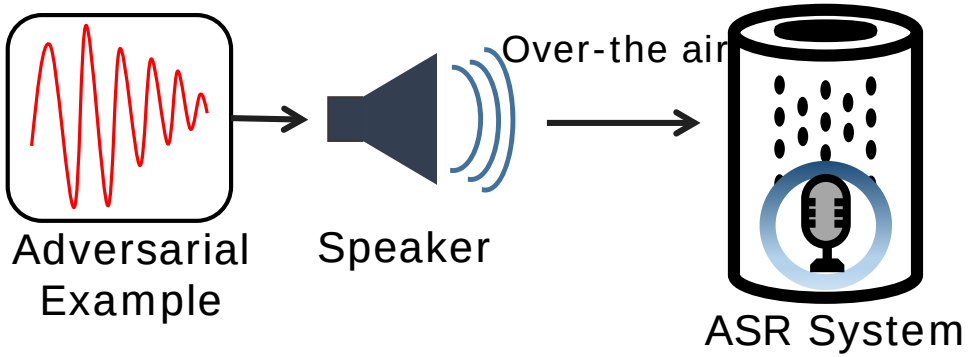
Evaluated Attack Types

- Two types of attack
  - Over-line attack



Target: open the door

  - Over-air attack



Target: open the door

Experimental Setup

- Target ASR system: DeepSpeech
- Speaker: DELL N889
- Receiver: Samsung Galaxy S8

- 3 over-line attacks and 2 over-air attacks
  - Carlini (S&P workshop 2018)
  - Taori (S&P workshop 2019)
  - W-S (AAAI 2020)
  - Hiromu (IJCAI 2019)
  - Metamorph (NDSS 2020)

Evaluation against over-line Attack

- Carlini / Taori : 500 benign samples, 500 adversarial examples
- W-S : 11 benign samples, 6 adversarial examples

		R.N.P(%)	TD1(%)	TD2(%)	Reverb(%)	Ours(%)
Carlini	FPR	**0**	0.6	0.6	**0**	**0**
	FNR	**0**	7.8	7.8	**0**	**0**
Taori	FPR	0	23.0	20.0	**0**	**0**
	FNR	0.4	7.6	2.2	**0**	**0**
W-S	FPR	**0**	**0**	0	**0**	**0**
	FNR	**0**	**0**	18.2	**0**	**0**

Evaluation against Over-air Attack

- Hiromu : 500 benign samples, 500 adversarial examples
- Metamorph : 4 benign samples, 16 adversarial examples

		R.N.P(%)	TD1(%)	TD2(%)	Reverb(%)	Ours(%)
Hiromu	FPR	0	5.4	0.6	2.0	**0.6**
	FNR	98.6	58.8	61.2	10.6	**1.4**
Metamorph	FPR	0	0	0	0	**0**
	FNR	100	68.8	56.3	43.8	**6.3**

Conclusion

- We make a robust detection system by analyzing logit gap difference
- The proposed method has much better performance in detecting over-air attacks.

Question?

Thank you!

Namgyu Park
Email: namgyu.park@postech.ac.kr

High Performance Computing Lab (**HPC**)
Pohang University of Science and Technology, South Korea
POSTECH