Learning Normality is Enough: A Software-based Mitigation against the Inaudible Voice Attacks

Xinfeng Li, Xiaoyu Ji*, Chen Yan, Chaohao Li, Yichen Li, Zhenning Zhang, Wenyuan Xu*

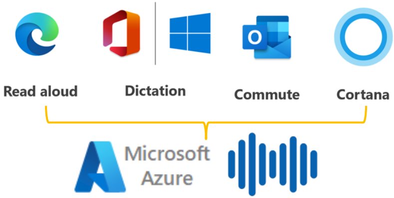
Ubiquitous System Security Lab (**USSLAB**), Zhejiang University

Voice Assistant Services are Everywhere!

Apple Siri

Microsoft Azure



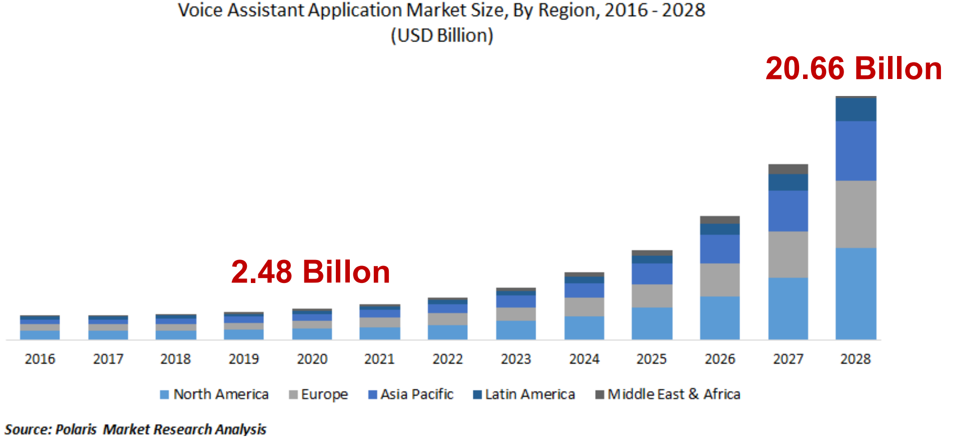
Google Home

Amazon Echo

Read my message

Call my boss

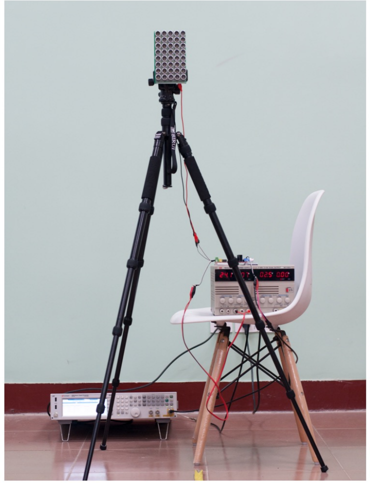
Open the door



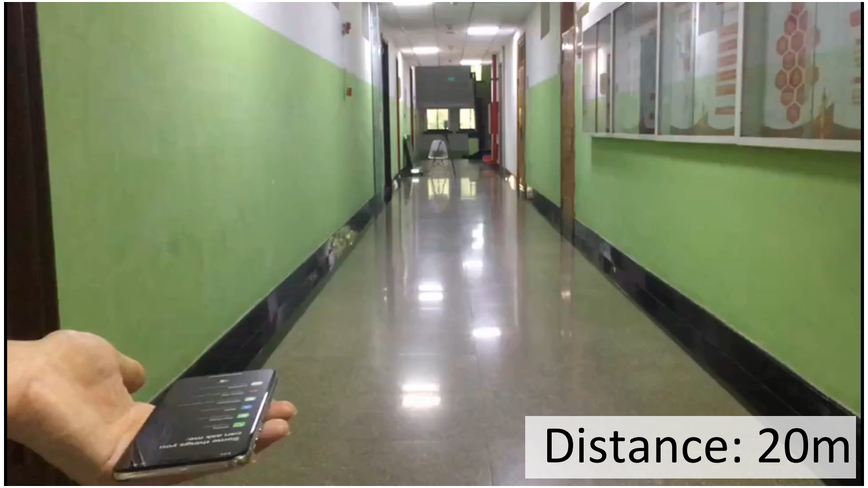
Inaudible Voice Attack (e.g., DolphinAttack)

- **Secretly injects malicious commands**
- **Inaudible to human beings**

DolphinAttack



Attack Device Setup

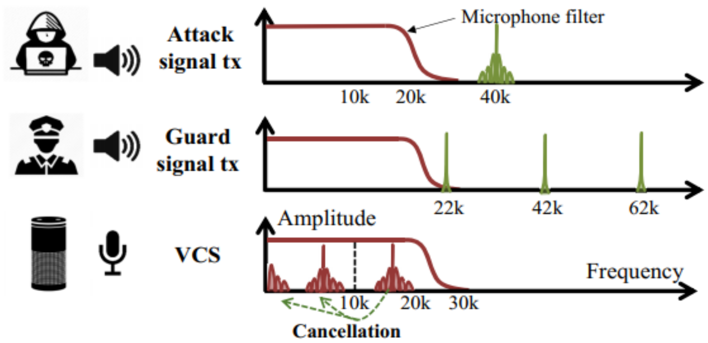


Real-world DolphinAttack: Control Siri from 20m away

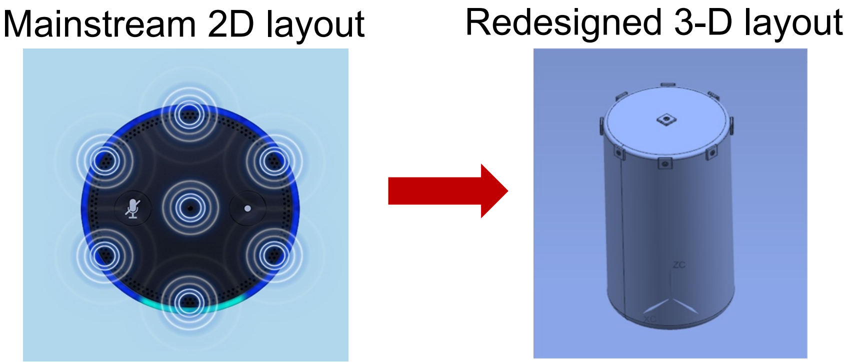
[1] Zhang, Guoming, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. "Dolphinattack: Inaudible voice commands." In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pp. 103-117. 2017.

Prior Defenses against Inaudible Voice Attacks

1. Hardware modification-based method He[MobiCom'20], EarArray [NDSS'21]



GuardSignal: actively emitting ultrasound

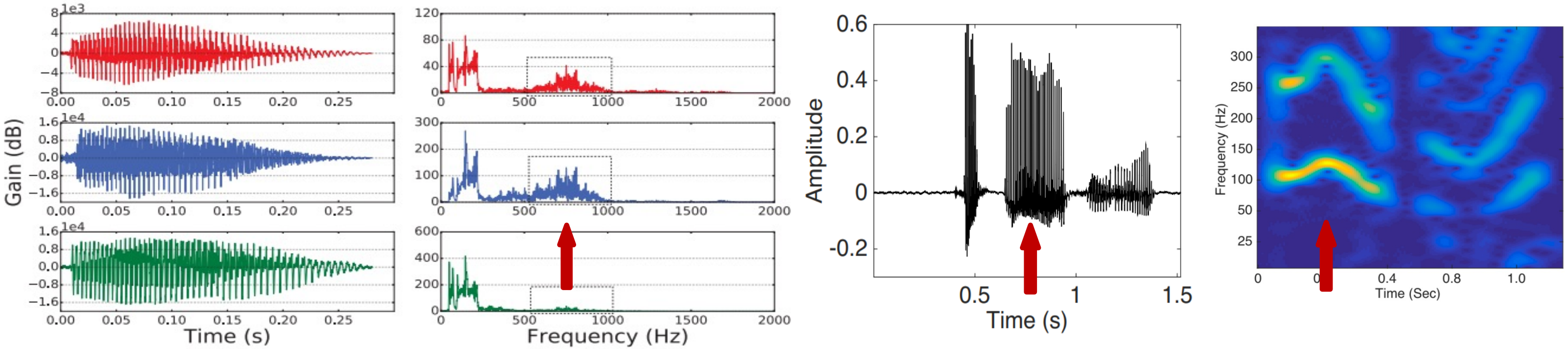


EarArray: re-design the microphone layout

Require hardware modification &
Cannot apply to billions of legacy devices

Prior Defenses against Inaudible Voice Attacks

2. Feature-based method leverages the traces of nonlinear effect and supervised classification Zhang [CCS’17], Roy [NSDI’18], Yan [NDSS’20], Li [CCS’21]



Supervised Classifiers Learn from Nonlinear Effect, e.g., Certain Frequencies / Signal Skewness of Attack Data

Nonlinear effects are device-dependent &
Require attack data collection and labeling

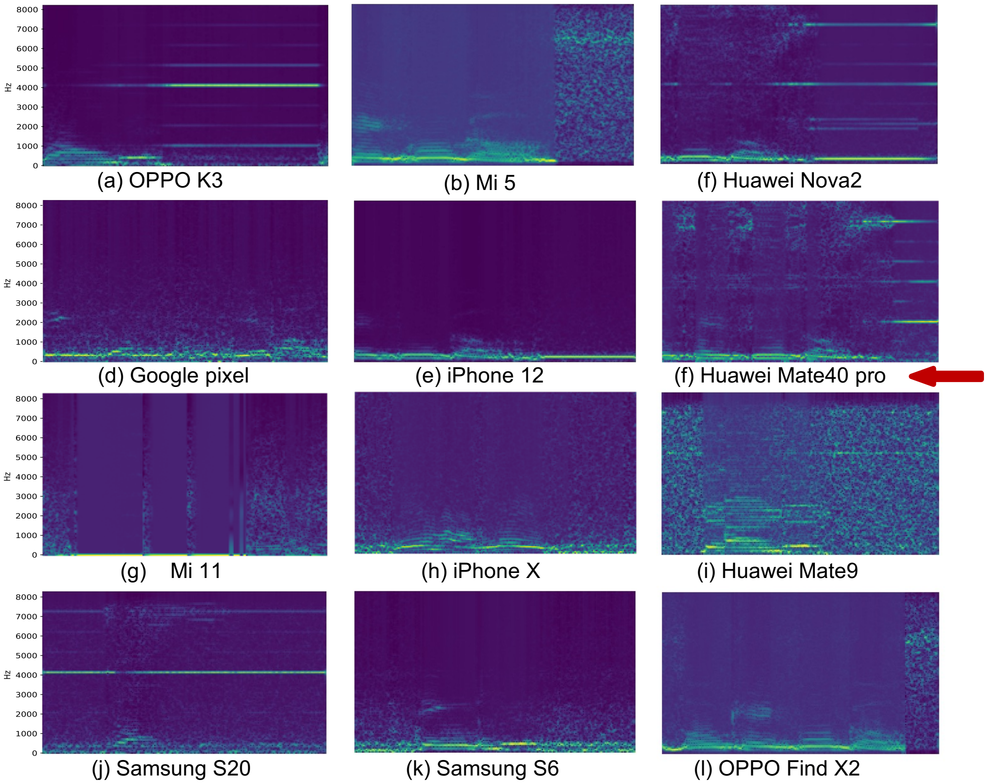
Prior Defense Limitations

1. **Hardware-based:** cannot apply to legacy devices

2. **Device-dependent:** cannot transfer to other devices

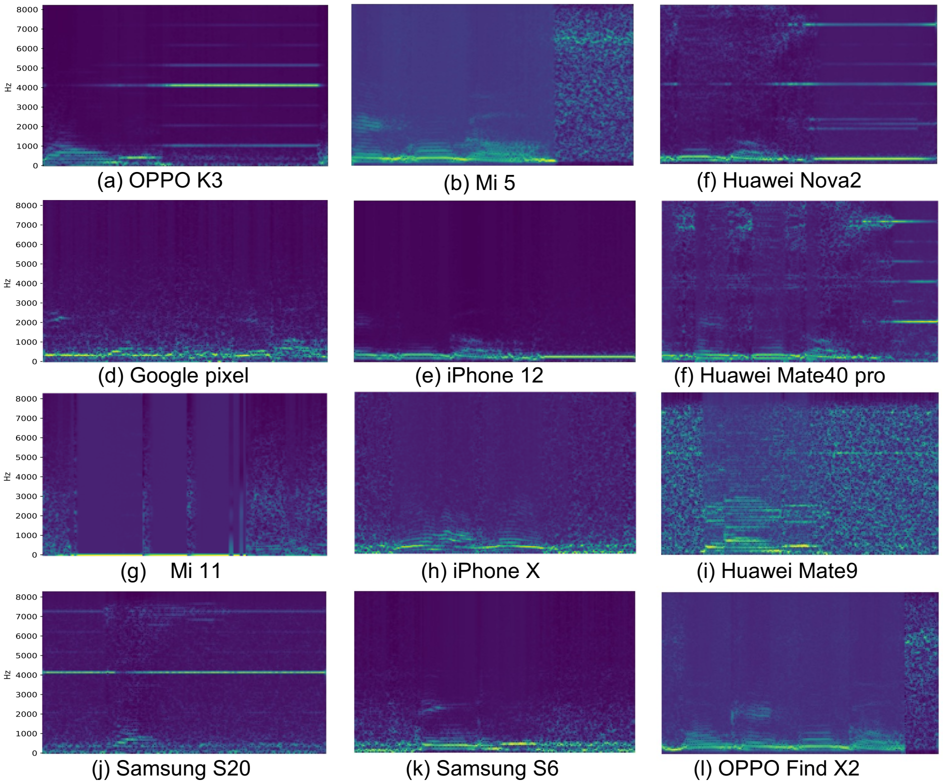
3. **Supervised:** require collecting lots of attack data

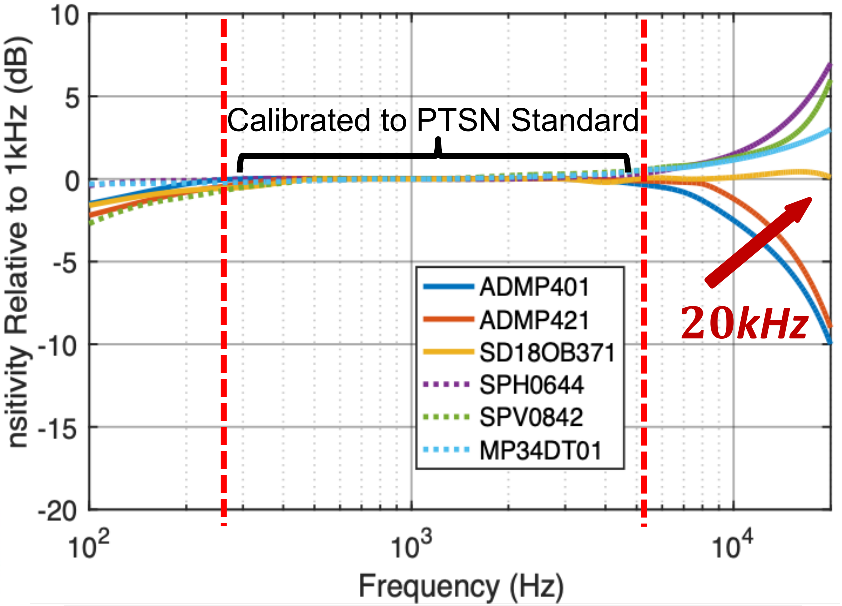
Spectra of inaudible commands on various devices



The same inaudible voice command ("OK Google") behaves differently on 24 devices

Spectra of inaudible commands on various devices

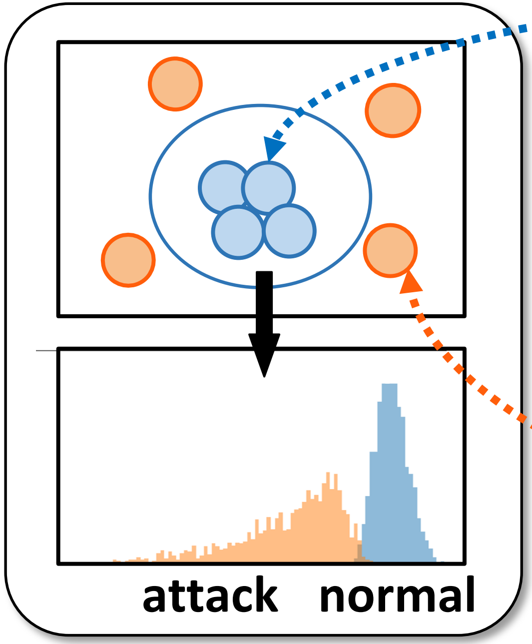


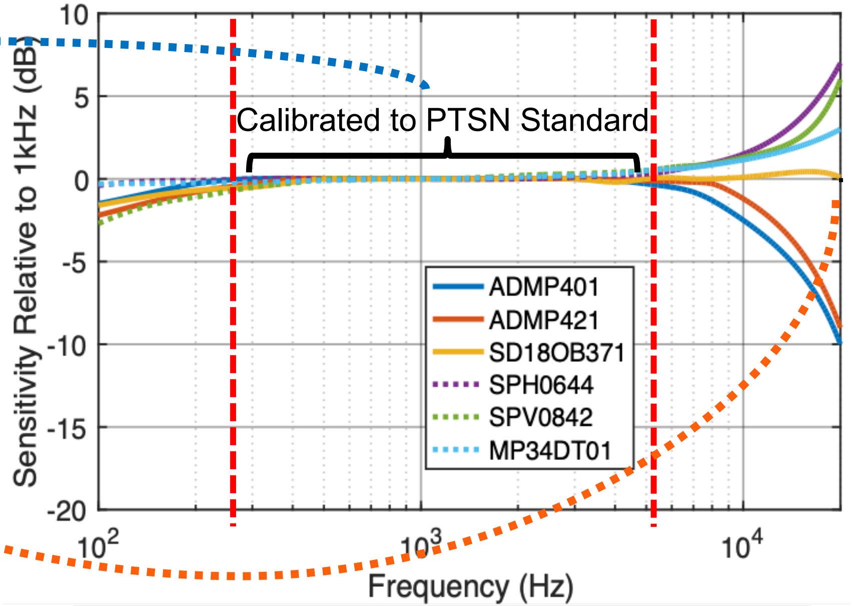


Different Microphones' Frequency Responses

Our Basic Idea

Normal (legal audio) Anomaly (attack audio)

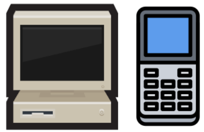




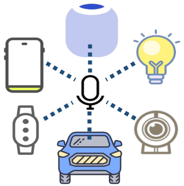
Different Microphones' Frequency Responses

NormDetect Wish List

1. Hardware-based → **1. Software-based:**
instantly protect legacy devices

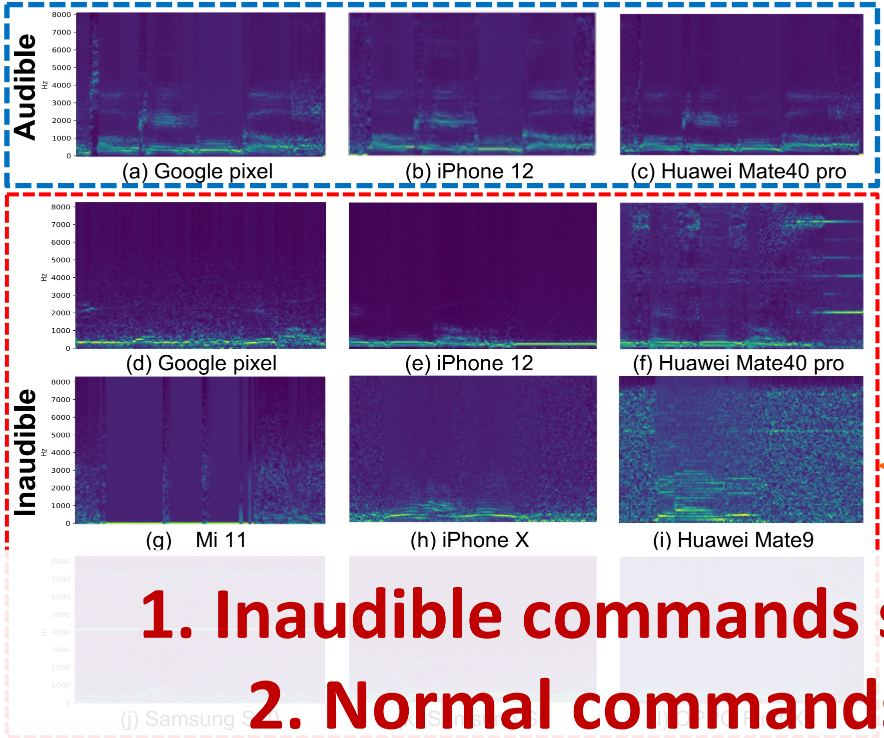


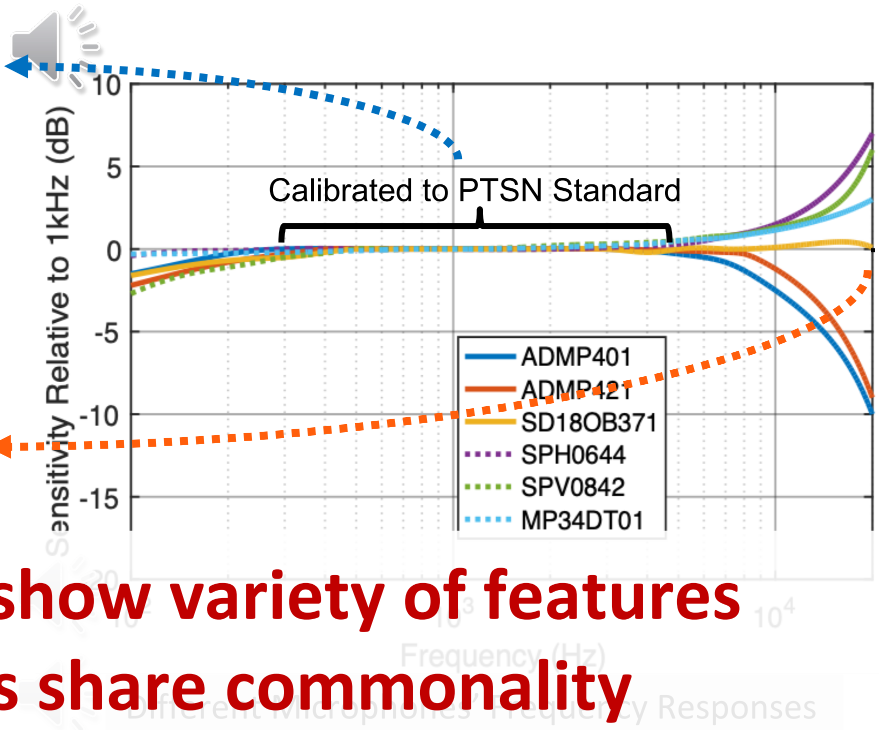
2. Device-dependent → **2. Universal:**
device-independent



3. Supervised → **3. Unsupervised:**
not require to collect attack data

Preliminary study for the command “OK Google”



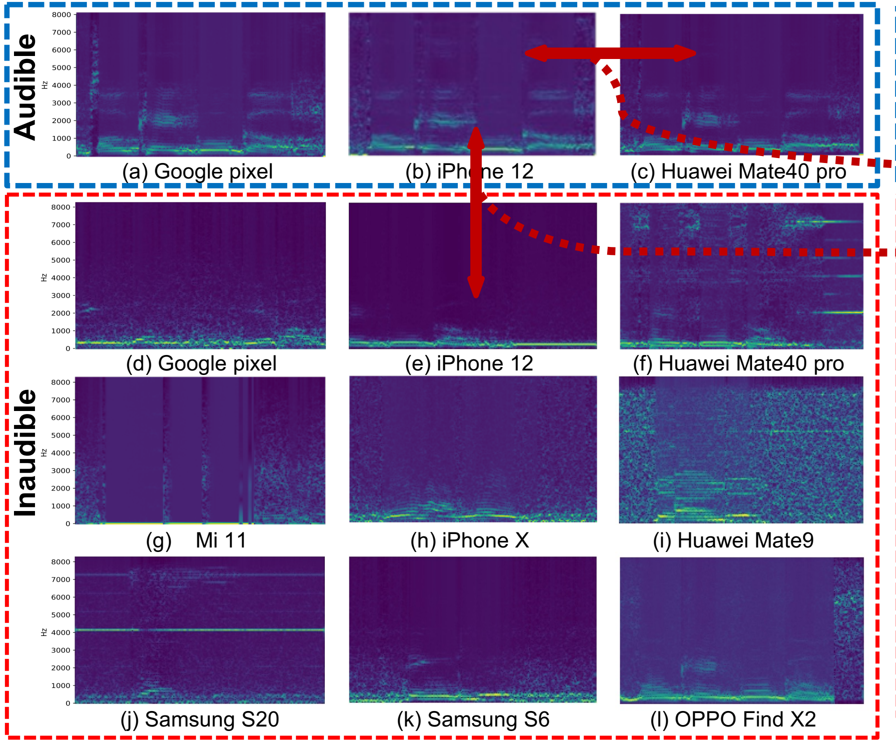


1. Inaudible commands show variety of features

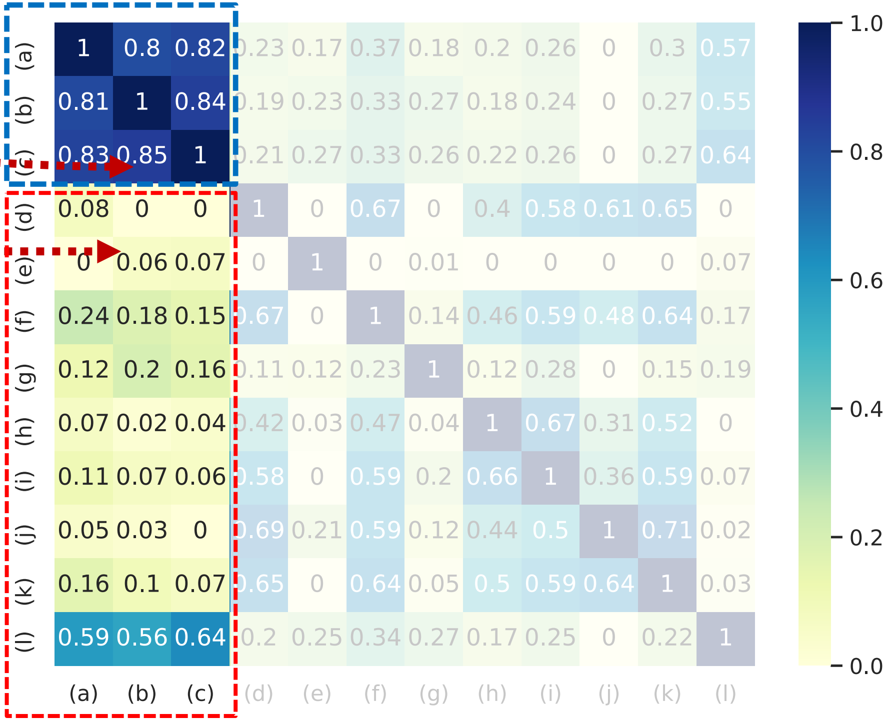
2. Normal commands share commonality

Audio spectra of “OK Google” on different devices

Preliminary study for the command “OK Google”

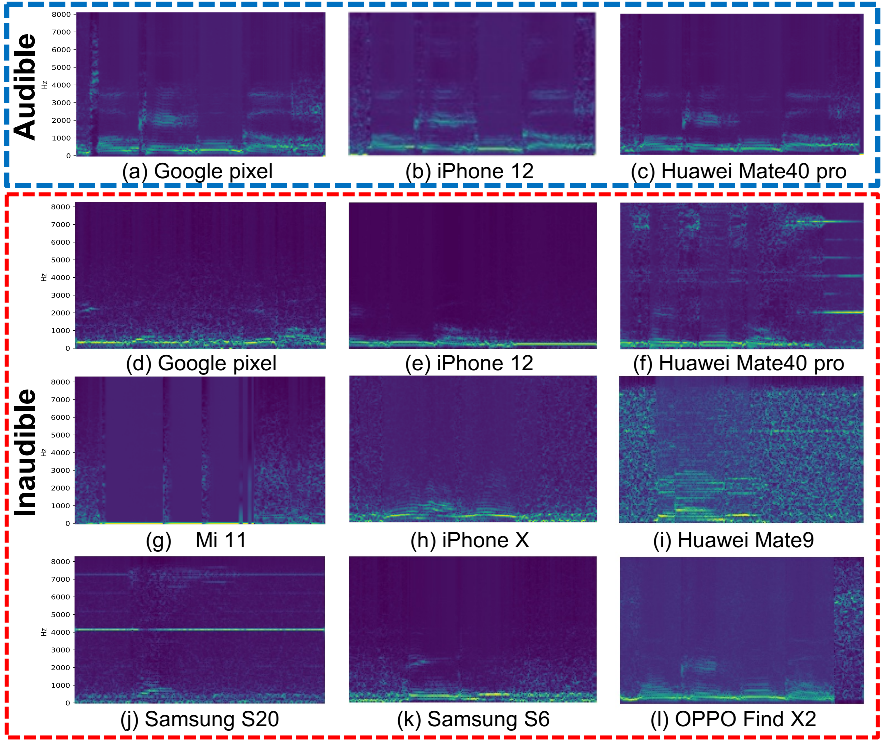


Audio spectra of “OK Google” on different devices

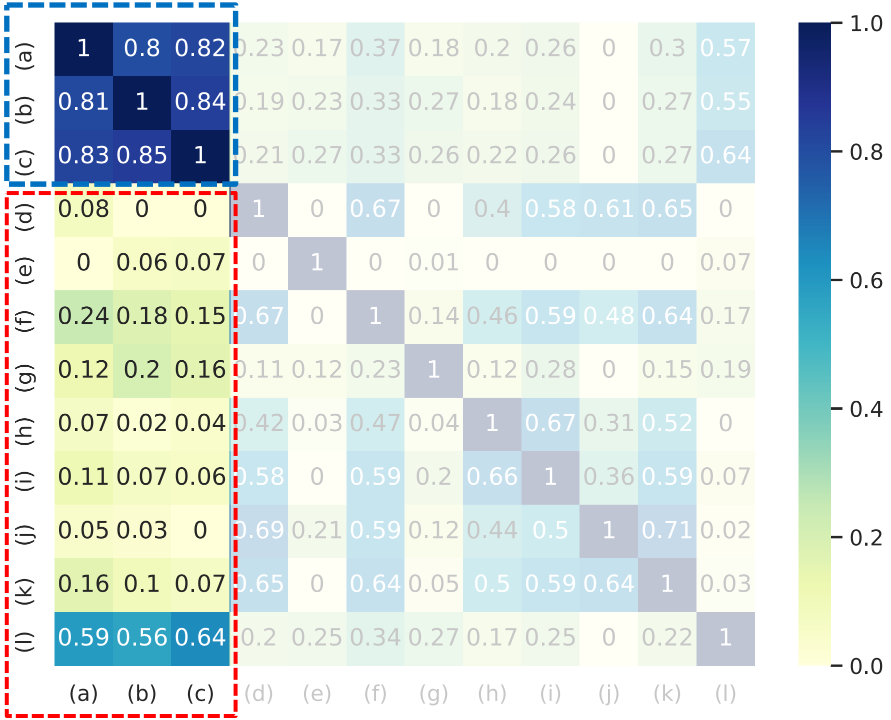


Heatmap of similarity scores

Preliminary study for the command “OK Google”

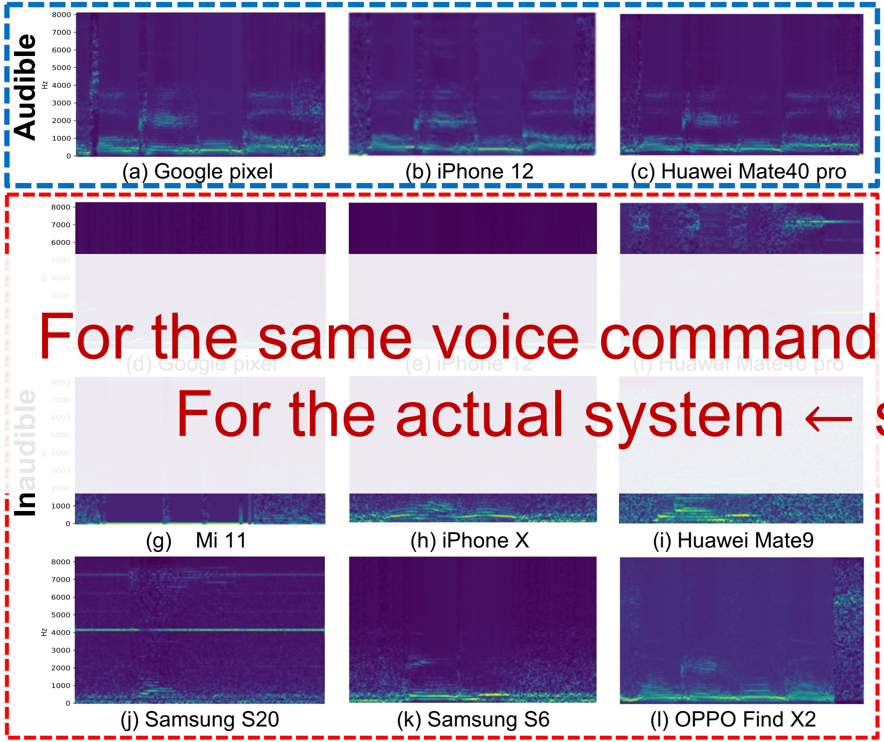


Audio spectra of “OK Google” on different devices

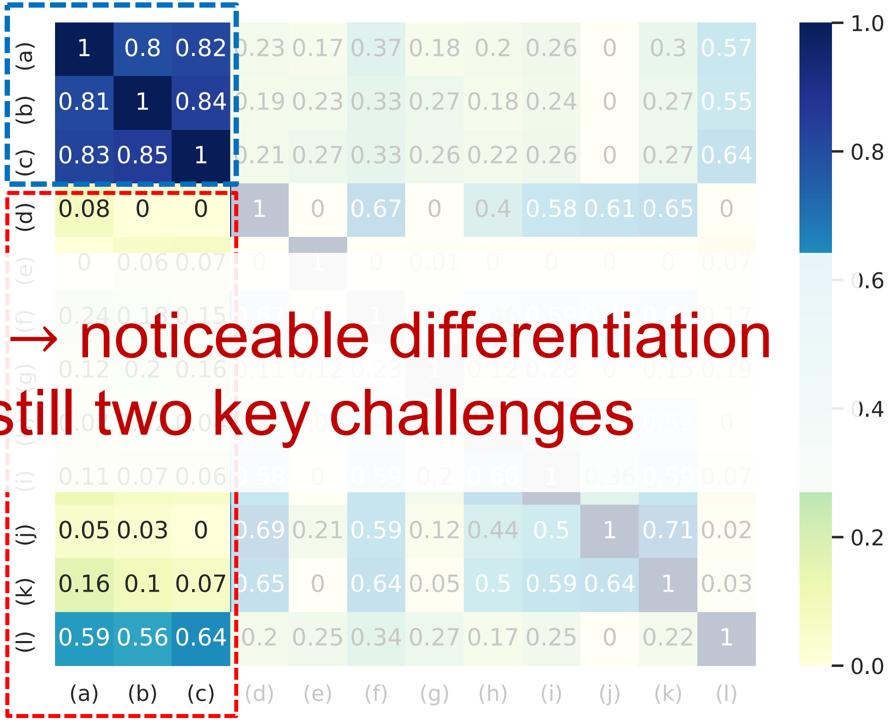


Heatmap of similarity scores

Preliminary study for the command “OK Google”



Audio spectra of “OK Google” on different devices



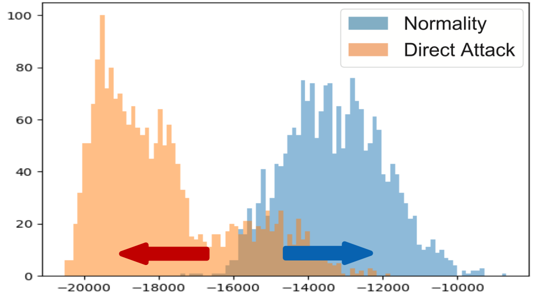
Heatmap of similarity scores

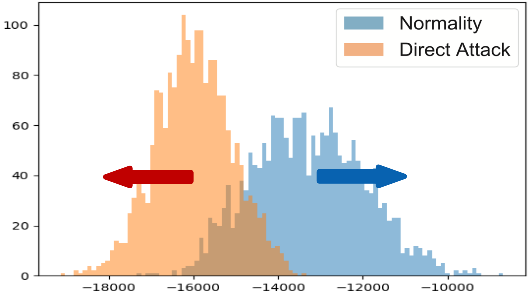
For the same voice command → noticeable differentiation

For the actual system ← still two key challenges

Key Challenges

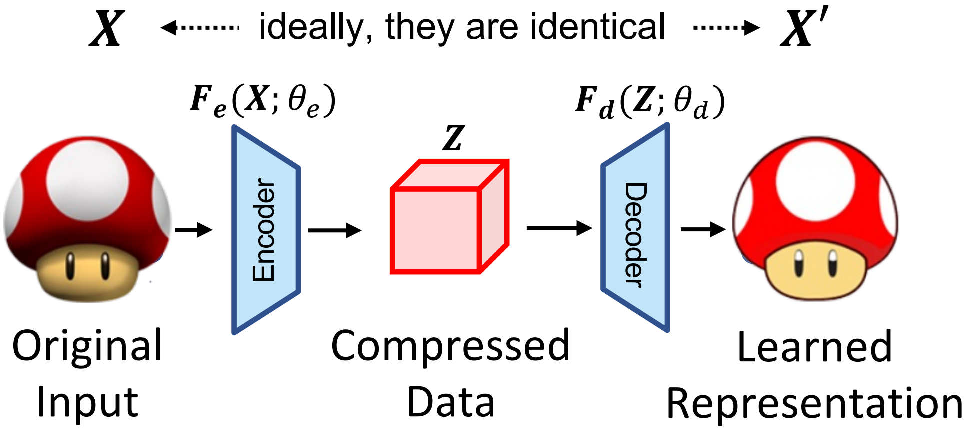
- ***Variance:*** *Normal audios may appear differently due to ambient noise, speakers, and voice content, etc.*
- ***Unsupervised****: How to reliably detect attacks without any attack data as prior knowledge for training.*





Reconstruction score distributions between the normal and attack

NormDetect's Basic Idea



① **Compressed Data:**

$$Z = F_e(X;\theta_e)$$

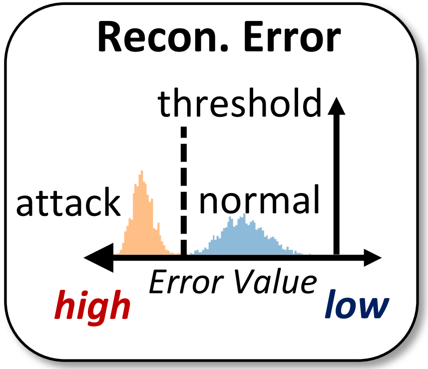
② **Learned Representation:**

$$X' = F_d(Z;\theta_d)$$

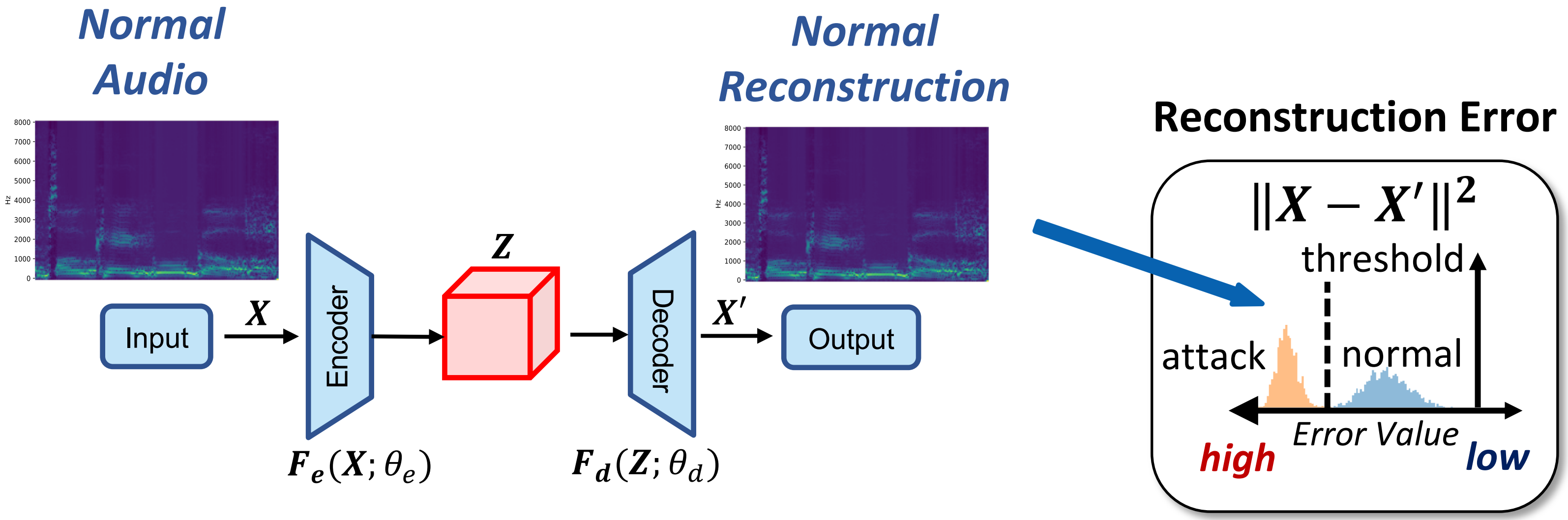
③ **Reconstruction Error:**

$$\|X - X'\|^2$$

Autoencoder: Anomaly Detection

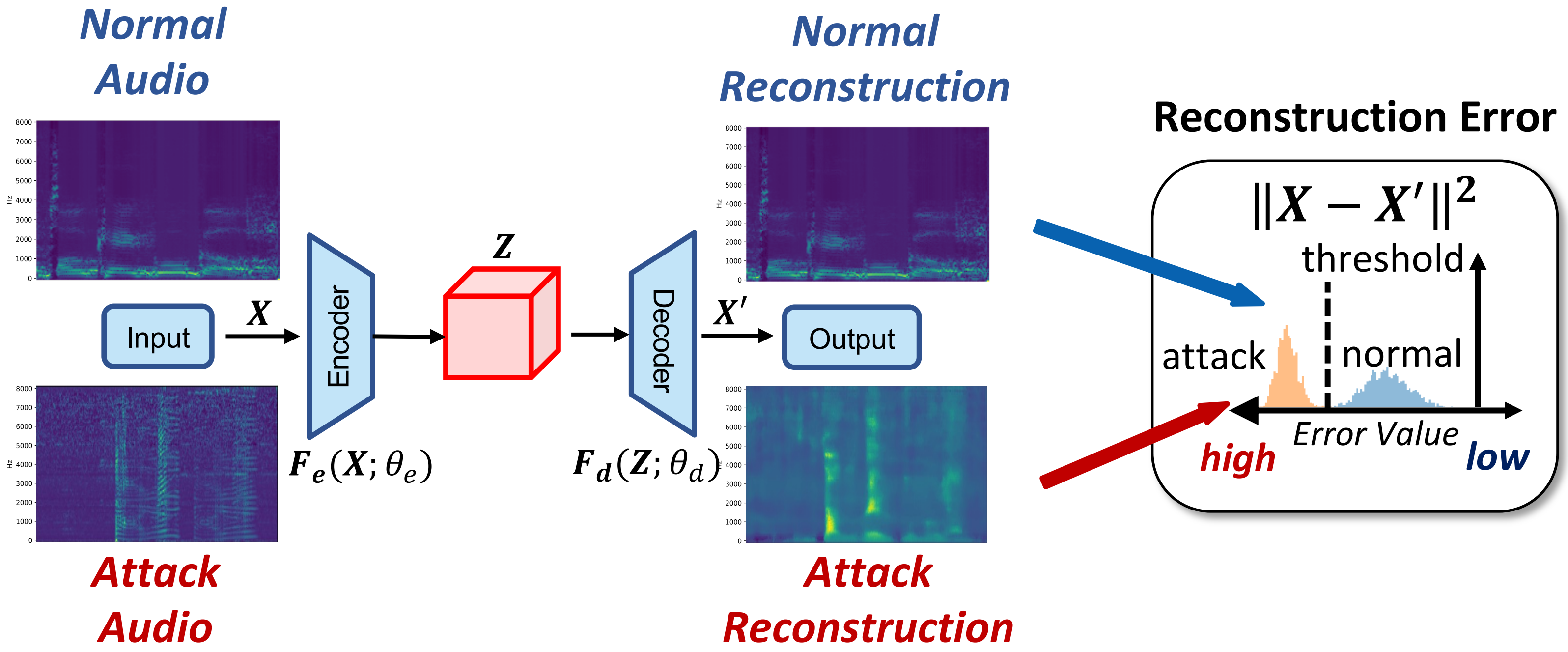


NormDetect's Basic Idea



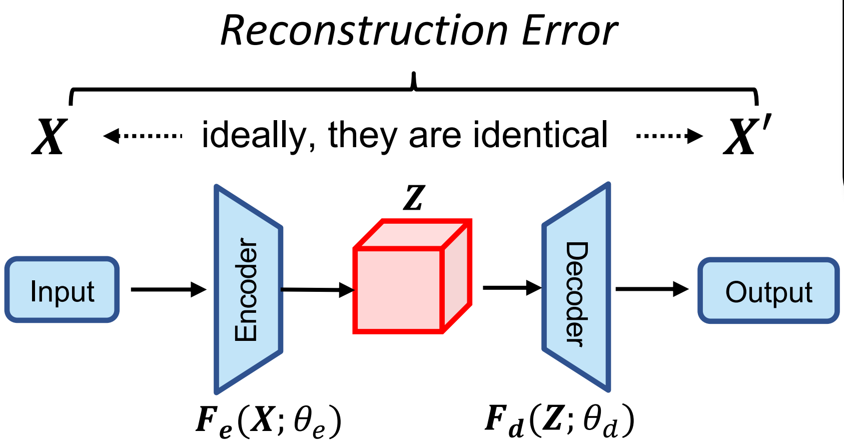
Autoencoder: Anomaly Detection

NormDetect’s Basic Idea

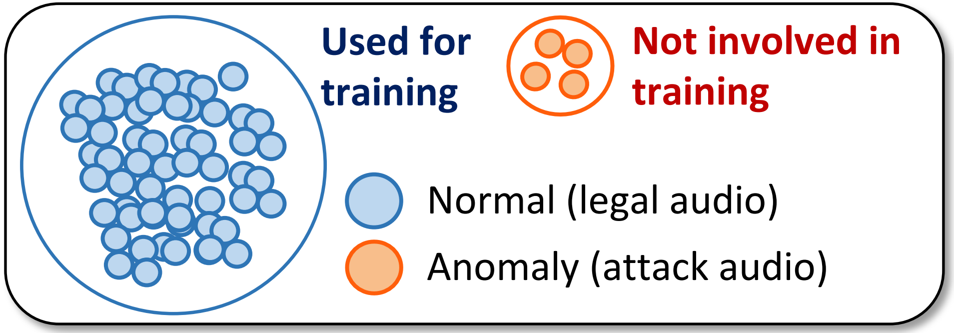


Autoencoder: Anomaly Detection

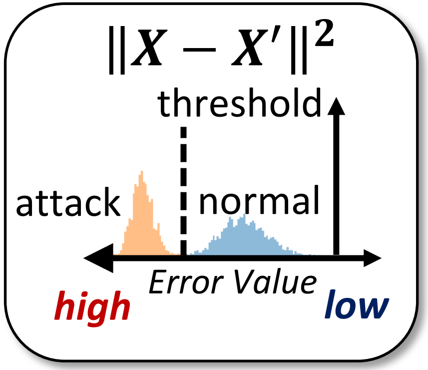
NormDetect's Basic Idea



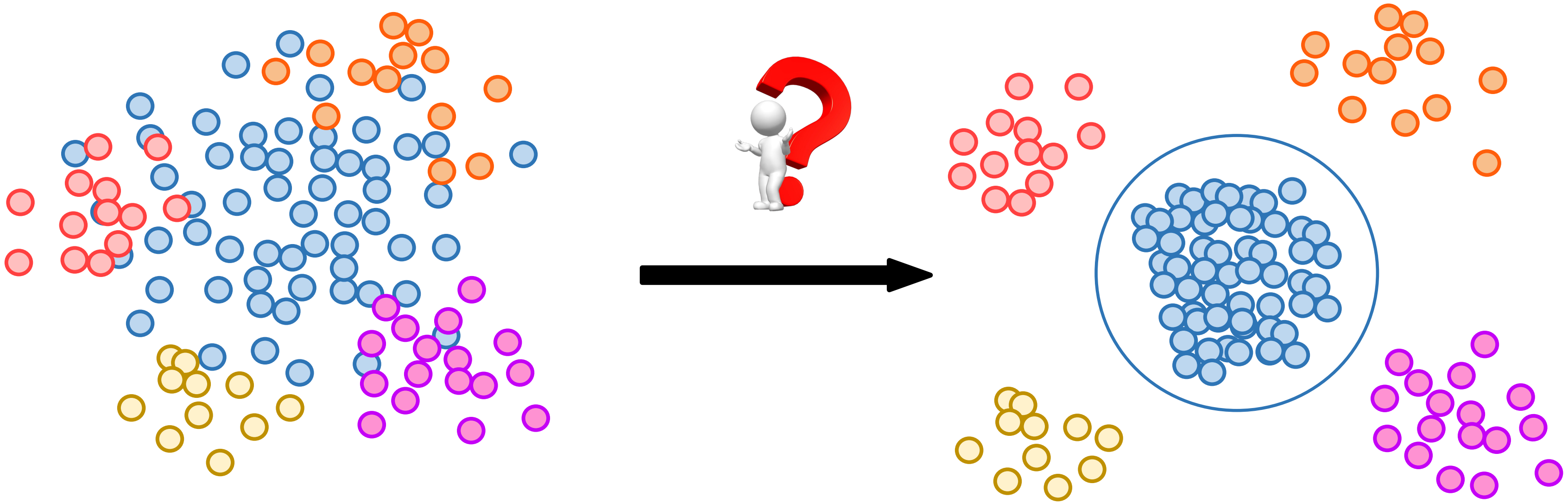
Autoencoder: Anomaly Detection



Reconstruction Error



Distribution is the Key Part



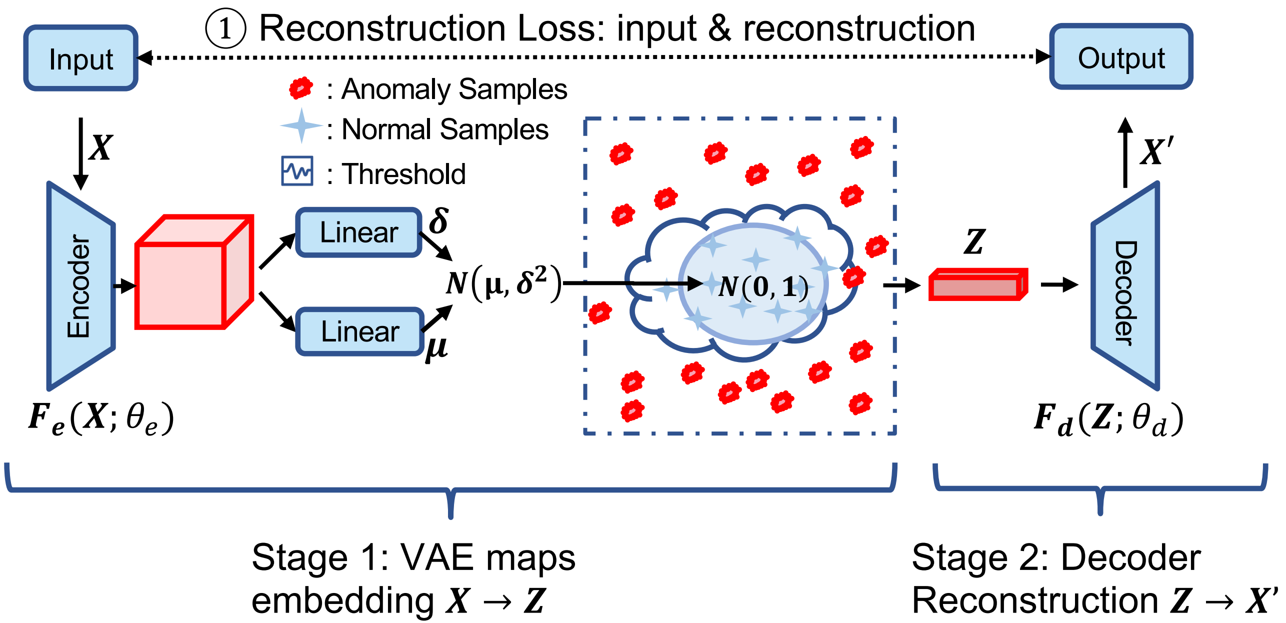
Loose and mixed representation distributions

Compact and separate representation distributions

: Normal patterns on different devices

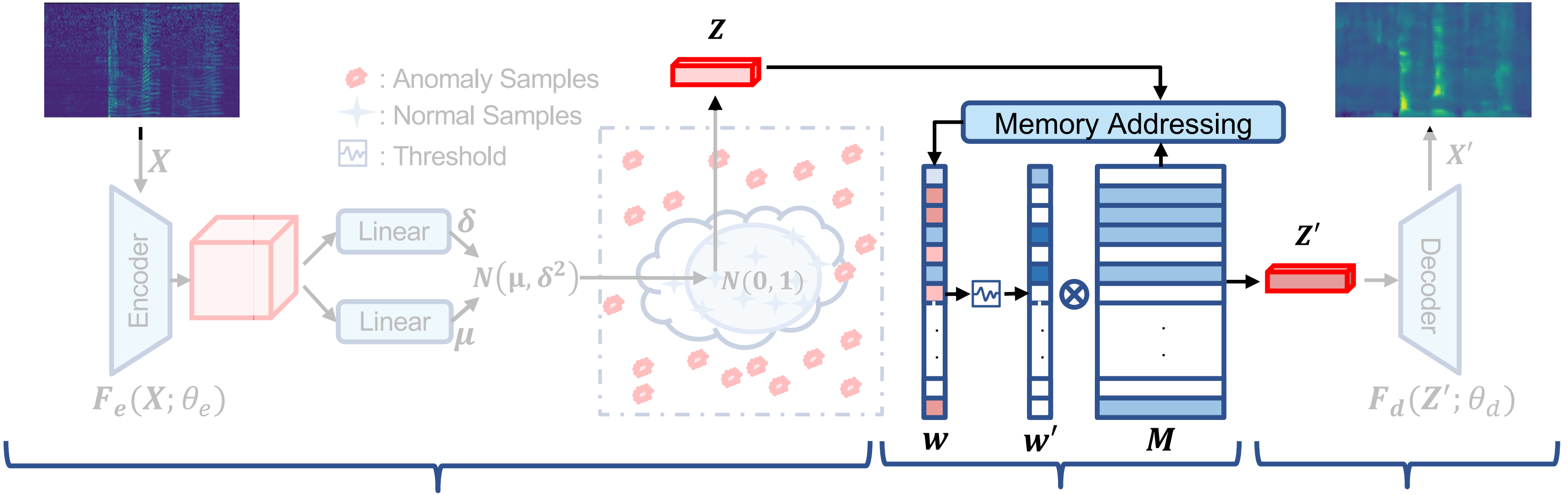
: Attack patterns on different devices

Distribution is the Key Part



① Reconstruction Loss **② Distribution Loss (KL divergence)**

Memorize the normality



Stage 1: VAE maps embedding $X \rightarrow Z$

Stage 2: Memory similarity query $Z \rightarrow Z'$

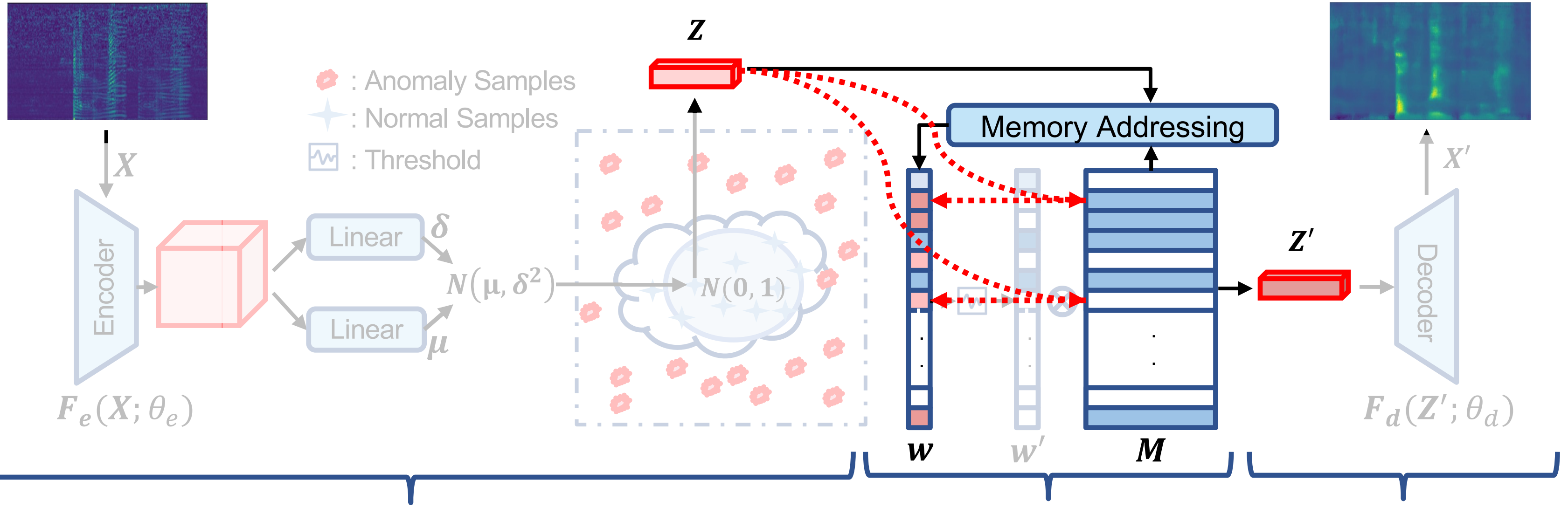
Stage 3: Decoder Reconstruction $Z' \rightarrow X'$

① Reconstruction Loss

② Distribution Loss (KL divergence)

③ Memory Similarity Query

Memorize the normality



Stage 1: VAE maps embedding $X \rightarrow Z$

Stage 2: Memory similarity query $Z \rightarrow Z'$

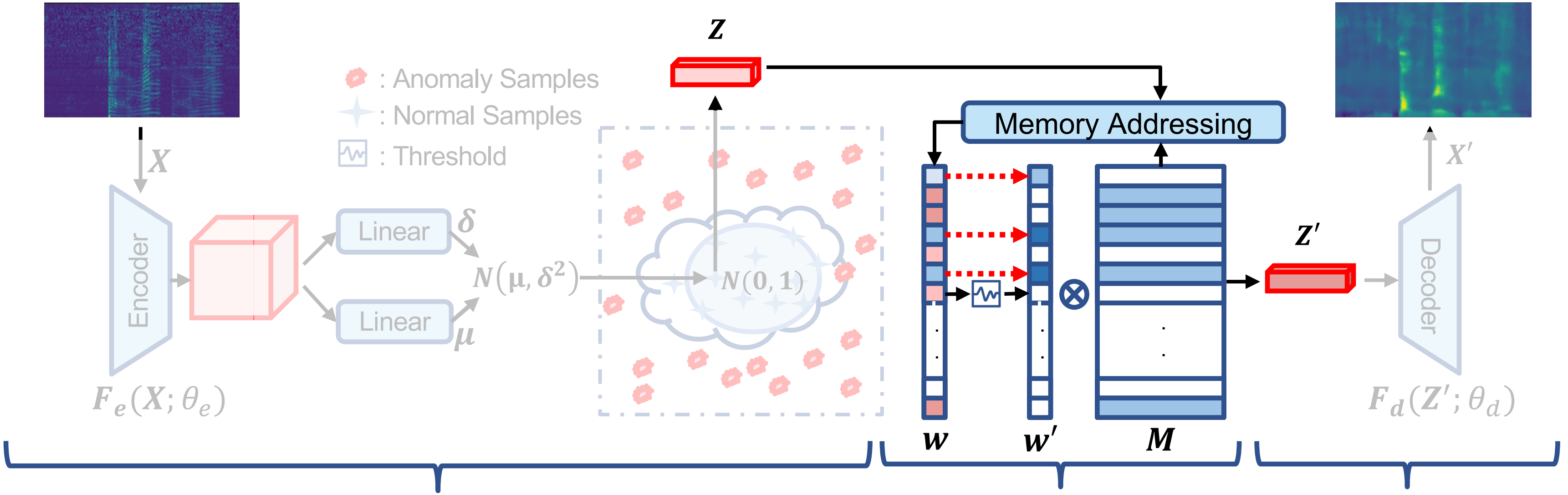
Stage 3: Decoder Reconstruction $Z' \rightarrow X'$

① Reconstruction Loss

② Distribution Loss (KL divergence)

③ Memory Similarity Query

Memorize the normality



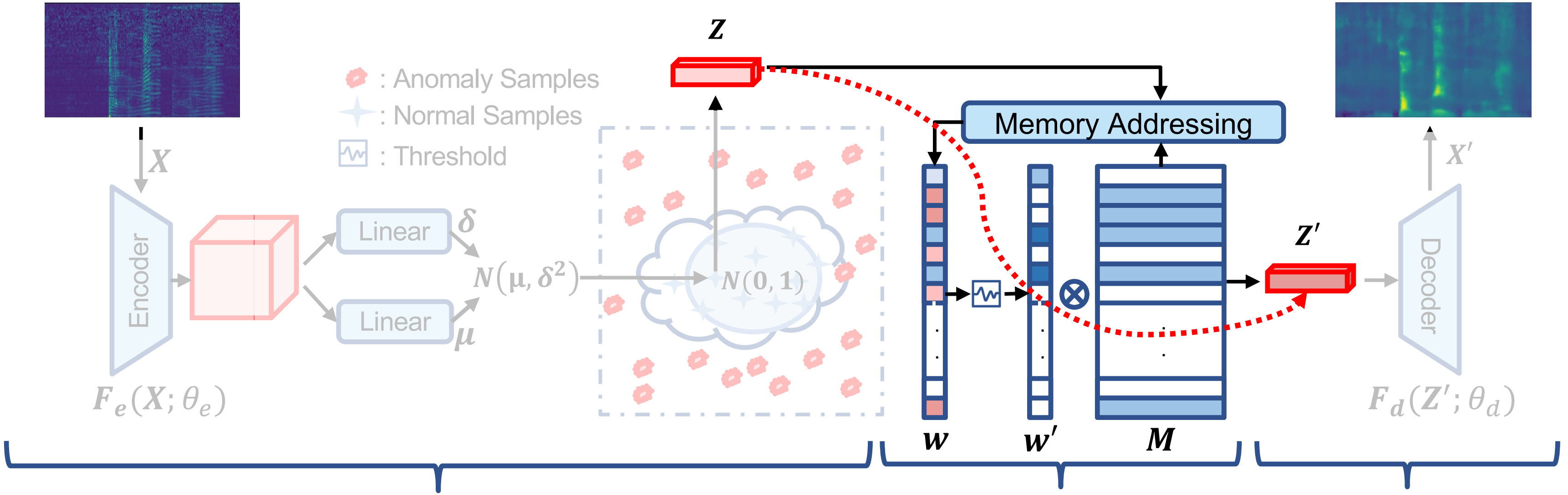
Stage 1: VAE maps embedding $X \rightarrow Z$

Stage 2: Memory similarity query $Z \rightarrow Z'$

Stage 3: Decoder Reconstruction $Z' \rightarrow X'$

① Reconstruction Loss **② Distribution Loss (KL divergence)** **③ Memory Similarity Query**

Memorize the normality



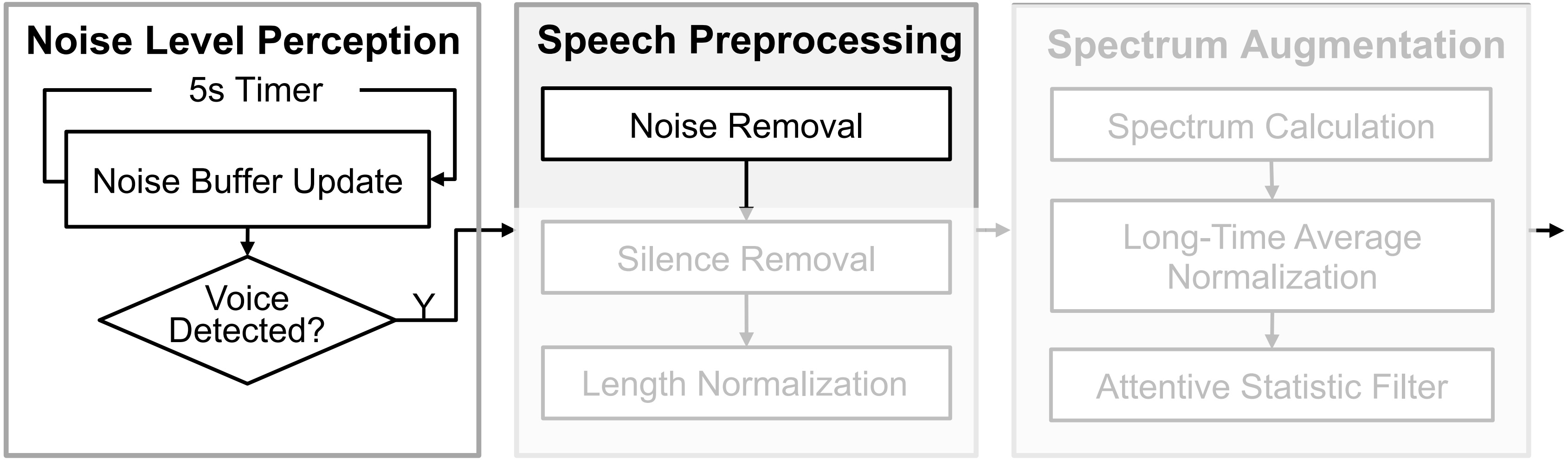
Stage 1: VAE maps embedding $X \rightarrow Z$

Stage 2: Memory similarity query $Z \rightarrow Z'$

Stage 3: Decoder Reconstruction $Z' \rightarrow X'$

① Reconstruction Loss **② Distribution Loss (KL divergence)** **③ Memory Similarity Query**

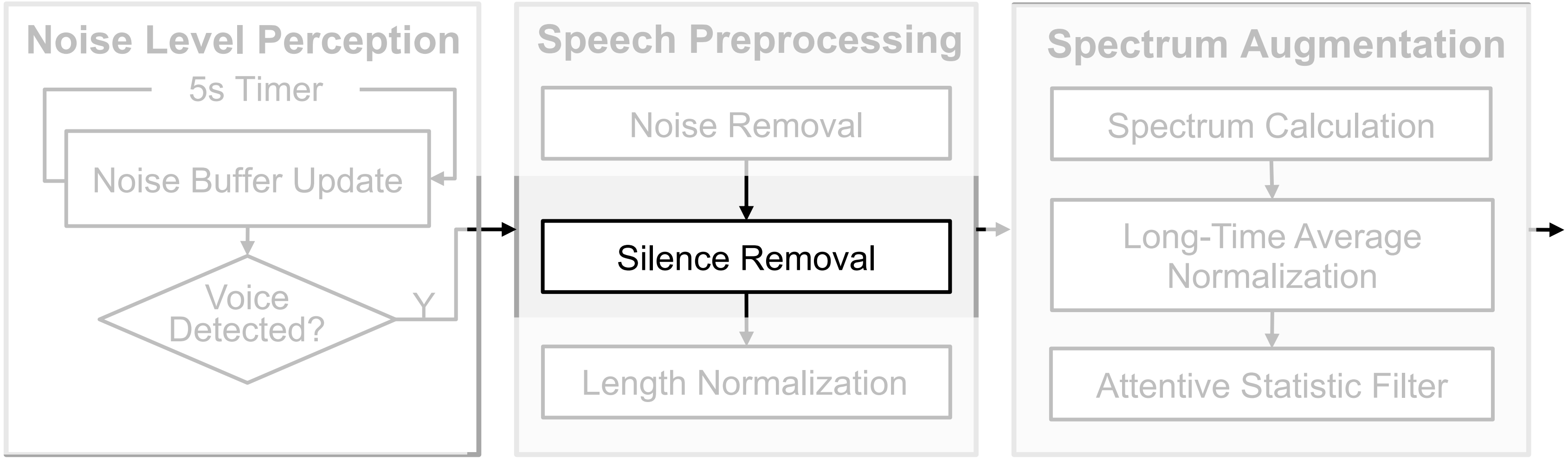
Reduce intra-normality variance



Ambient Noise

- periodic noise perception & removal
- different from attacks with anomalous noises

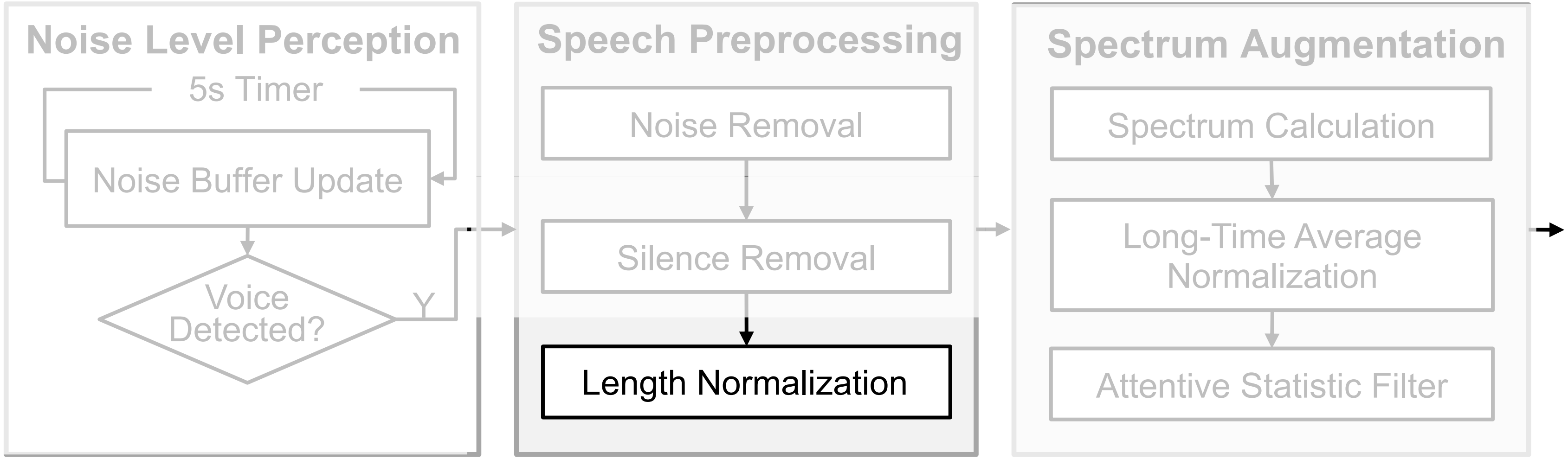
Reduce intra-normality variance

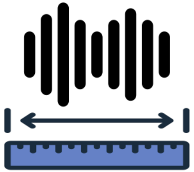


Speech Habit

- speech speed / semantic pause
- remove unnecessary silence clips

Reduce intra-normality variance

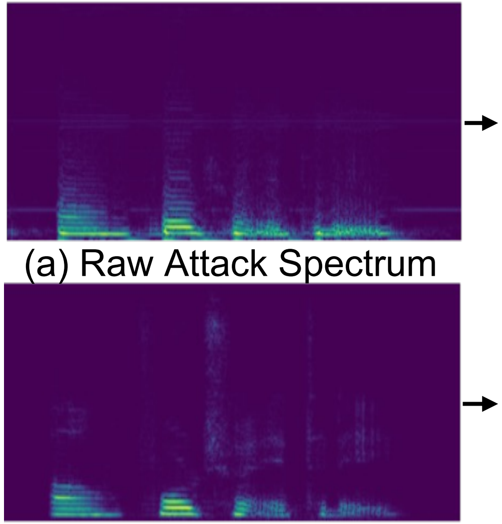




Speech Length

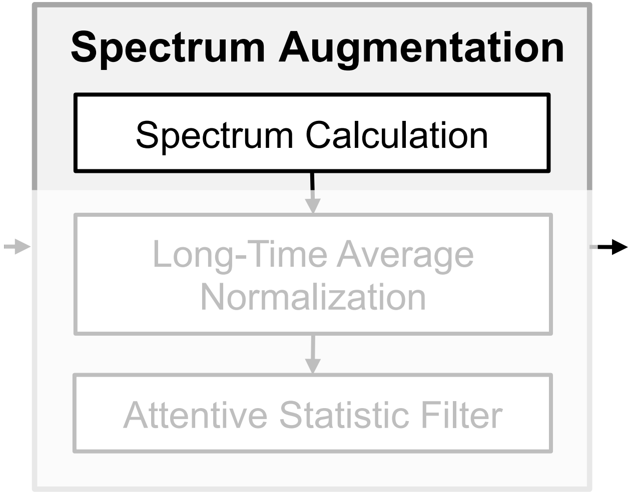
- different speech content length
- normalize to 1.5-second per frame

Increase Attack-Normal Differences

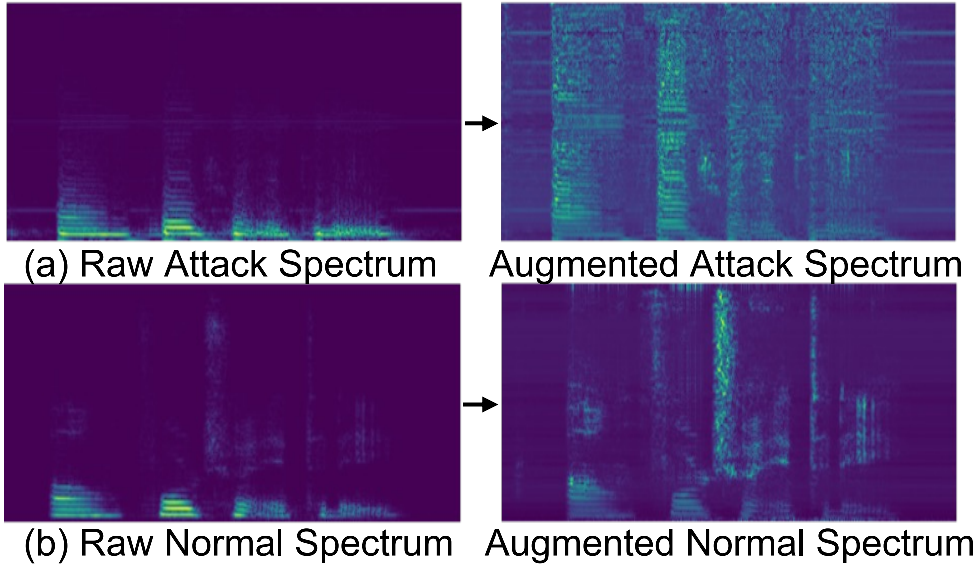


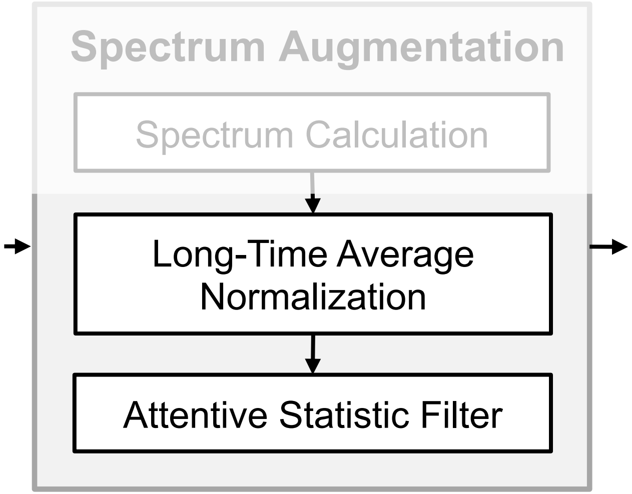
(a) Raw Attack Spectrum

(b) Raw Normal Spectrum



Increase Attack-Normal Differences





Evaluation

➢ **Training Dataset:**

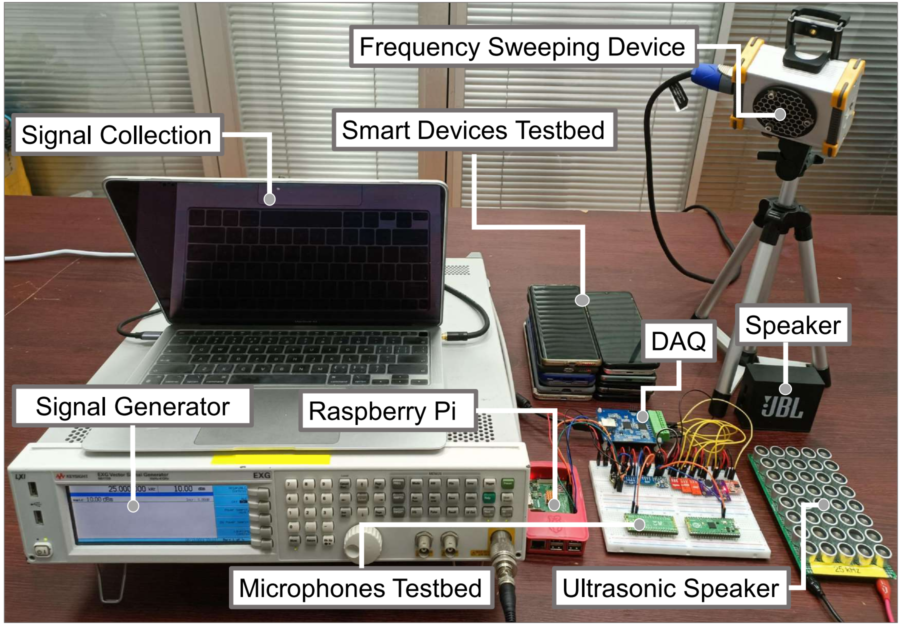
Fluent Speech Commands

- 30,042 pieces of English audio

➢ **Evaluation Dataset:**

Audible & Inaudible Voice Commands

- **7** Distances (10 ~ 300 cm)
- **24** mainstream devices (smartphone ~ smart watch)
- 28 speakers
- English & Chinese
- **383,320** pieces of audio



Experimental Setup

Evaluation

➢ **Training Dataset:**

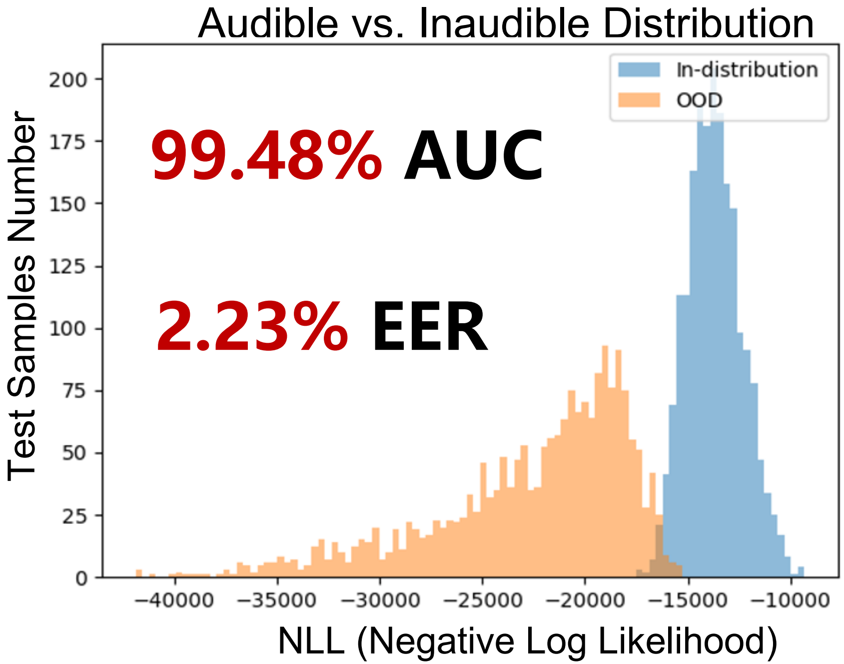
Fluent Speech Commands

- 30,042 pieces of English audio

➢ **Evaluation Dataset:**

Audible & Inaudible Voice Commands

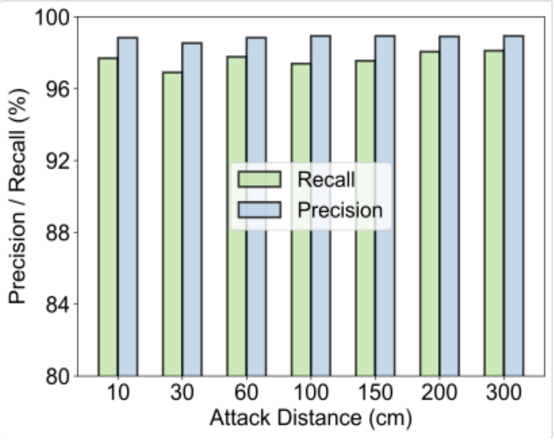
- **7** Distances (10 ~ 300 cm)
- **24** mainstream devices (smartphone ~ smart watch)
- 28 speakers
- English & Chinese
- **383,320** pieces of audio



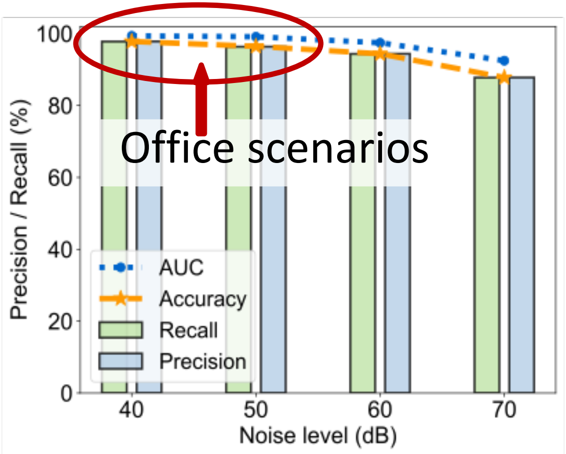
Evaluation

- ***A user is more concerned about NormDetect's effectiveness with:***

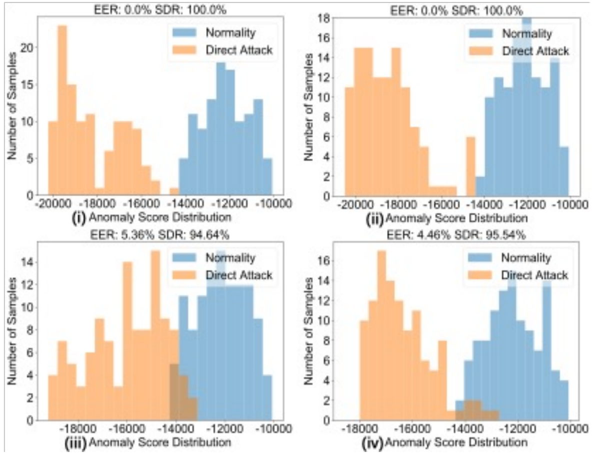
Attack distances



Ambient noise levels



Different device models



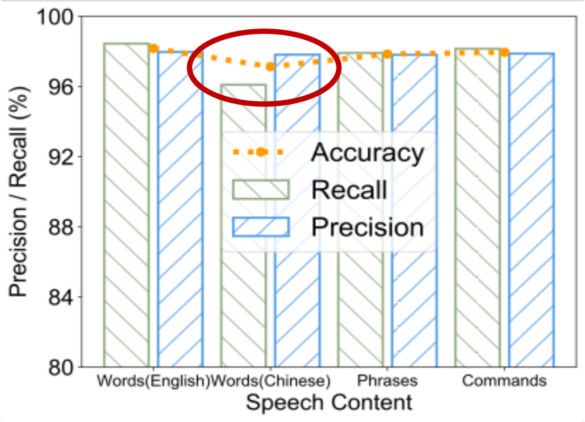
Precision/Recall keep >96% under most cases

SDR keep >94%

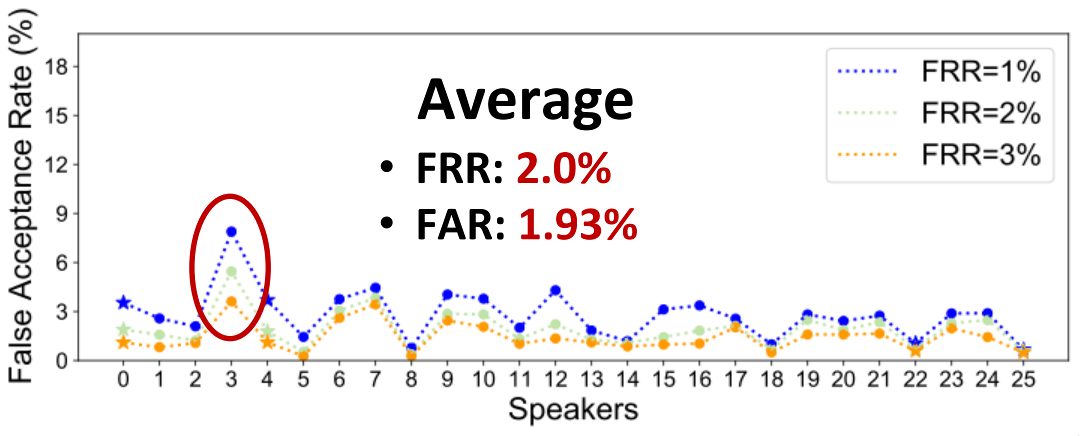
Evaluation

- *A user is also concerned about NormDetect's effectiveness with:*

Languages



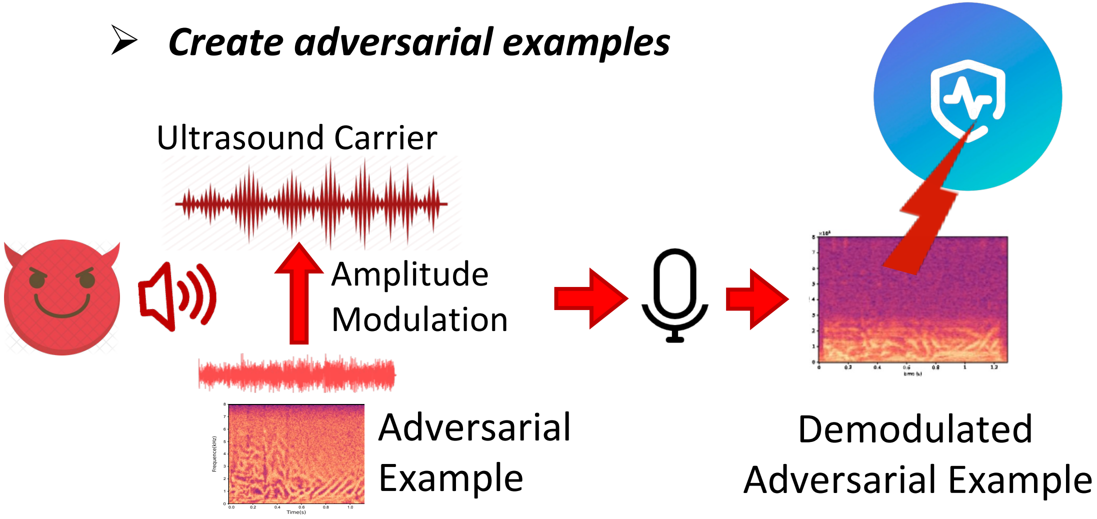
Speaker identities

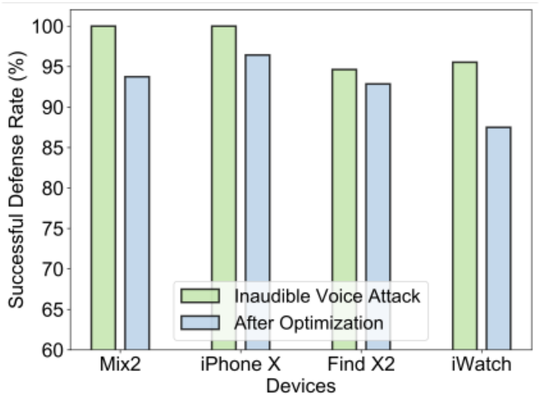


NormDetect can also adapt to unseen languages and speakers

Evaluation

- ***An adaptive adversary may try to:***
  - ***Create adversarial examples***





NormDetect maintain average SDR >92% under Adaptive Attacks

Conclusion

- First unsupervised software-based mitigation against the inaudible voice attacks.

- NormDetect is evaluated on the large audible & inaudible voice commands dataset consisting of 24 devices and 383,320 audios.

Learning Normality is Enough: A Software-based Mitigation against the Inaudible Voice Attacks

Contact the authors at:

xinfengli@zju.edu.cn

xji@zju.edu.cn

yanchen@zju.edu.cn

wyxu@zju.edu.cn

USSLAB

Homepage: www.usslab.org