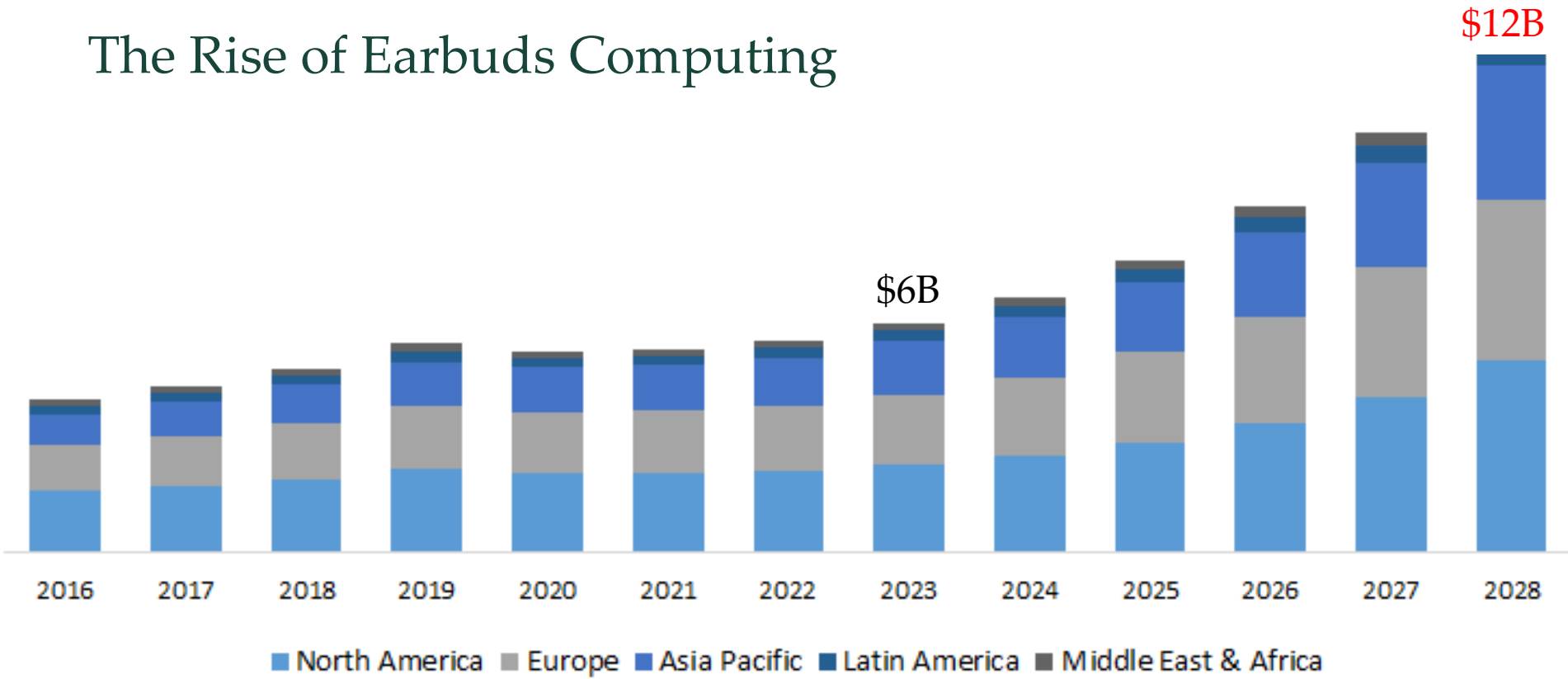
Towards Practical and Secure Earable Computing Using Multi-modal Inference

Tianxing Li
Department of Computer Science and Engineering
Michigan State University

June 3rd, 2024

The Rise of Earbuds Computing



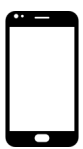
Source: Palaris Market Research Analysis

The Rise of Earbuds Computing

Wear earbuds on bus

Wear earbuds in gym

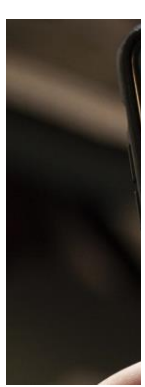
Handsfree Voice Controllable System

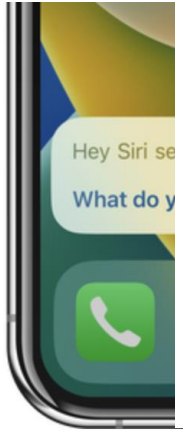


"play some music"

"Send message to XX"

"Call XXX"





Hey Siri

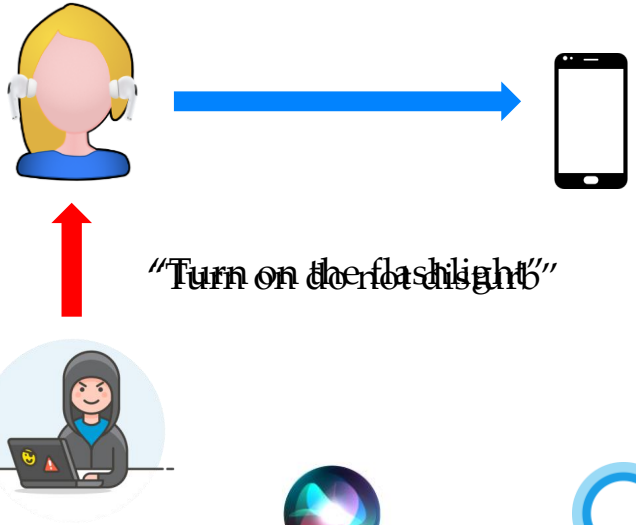
Hey Cortana

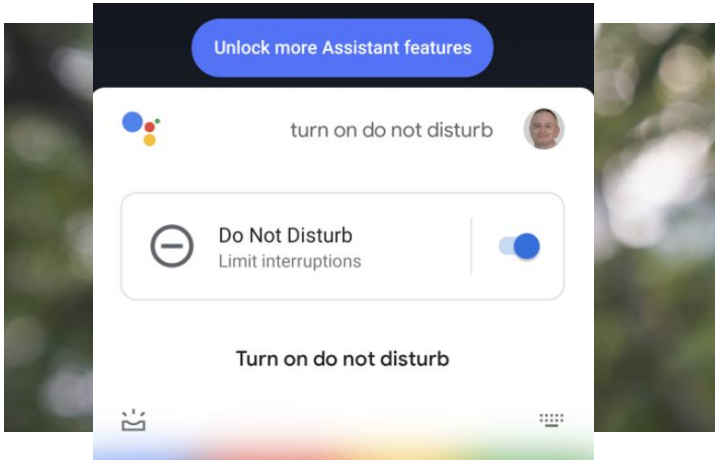
Alexa

Ok Google

Hi Bixby

Security Threats of Voice Controllable Systems





Hey Siri

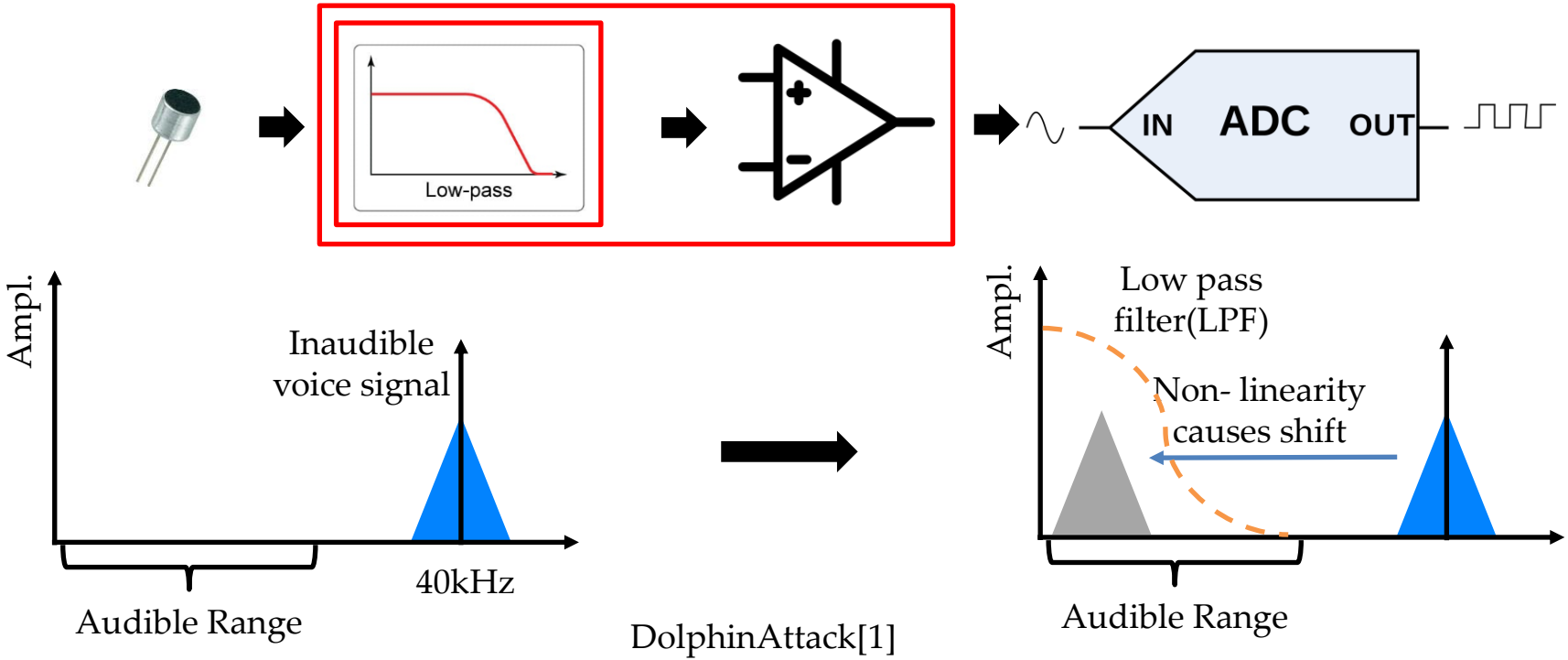
Hey Cortana

Alexa

Ok Google

Hi Bixby

Background: Non-linearity Effect of Microphone



[1] Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., & Xu, W. (2017, October). Dolphinattack: Inaudible voice commands. In Proceedings of the 2017 ACM SIGSAC conference on computer and communications security (pp. 103-117).

Are These Earbuds as Secure as Our Smartphones?

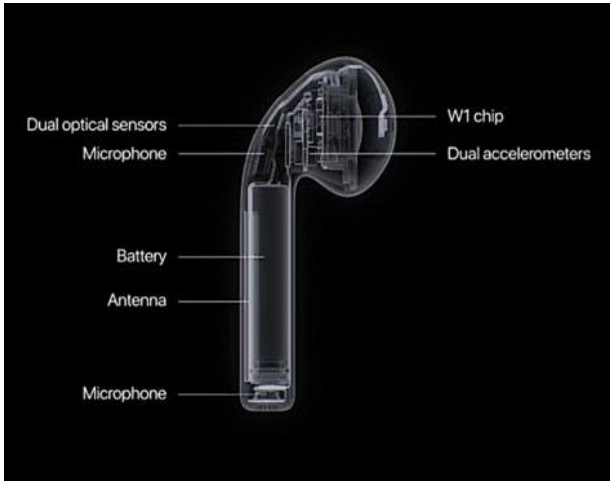
- LiVoAuth[1]
- CaField[2]
- VoiceGesture[3]

[1] Zhang, Rui, et al. "LiVoAuth: Liveness Detection in Voiceprint Authentication with Random Challenges and Detection Modes." *IEEE Transactions on Industrial Informatics* (2022).
[2] Yan, Chen, et al. "The catcher in the field: A fieldprint based spoofing detection for text-independent speaker verification." Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019.
[3] Zhang, Linghan, Sheng Tan, and Jie Yang. "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication." Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017.

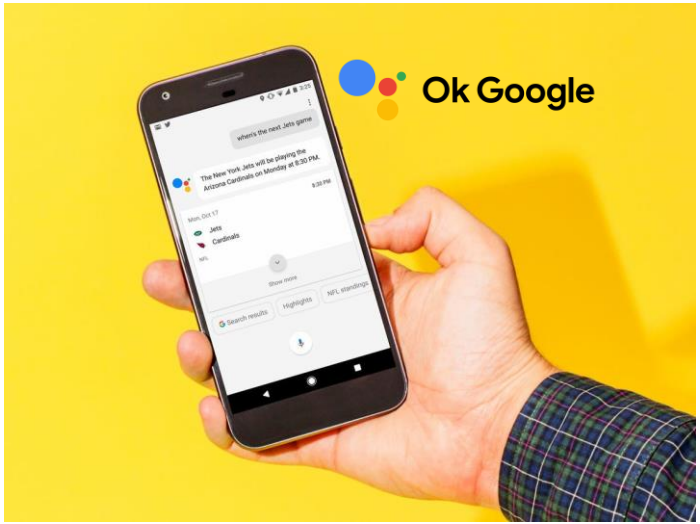
Challenges: Hardware Limitation.

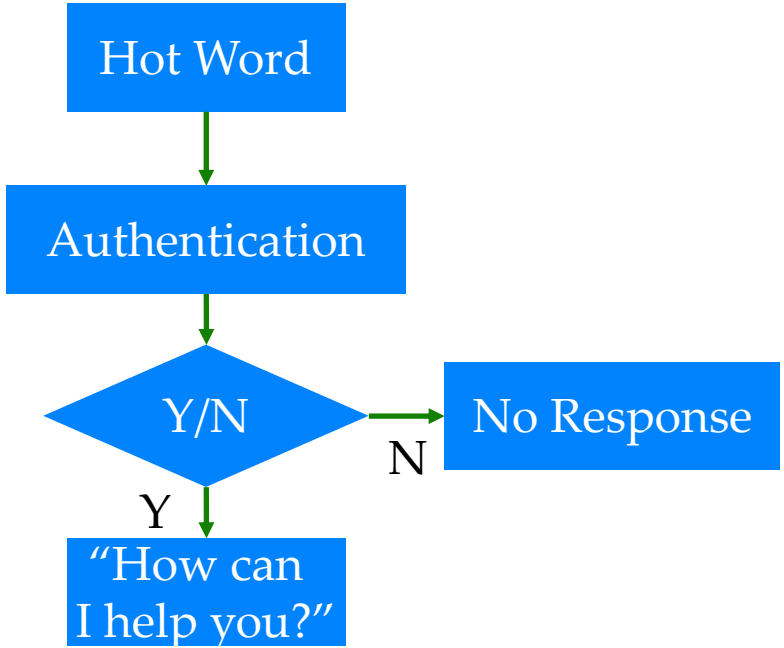
- Limited Computing Resources
  - H2 chip (Airpods 3) vs. A16 (iPhone 15)

- Limited Energy Budget
  - 369 mAh (Airpods 3) vs. 3367mAh (iPhone 15)



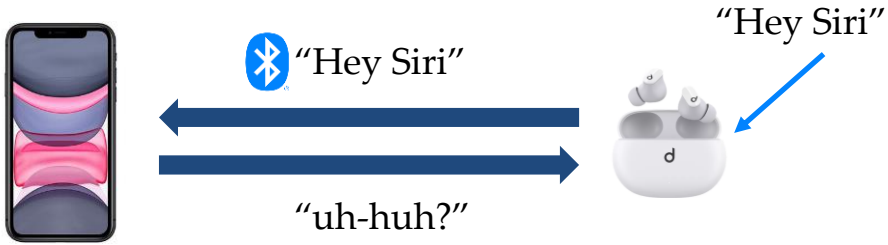
Software Vulnerability Study: Earbuds vs. Smartphone



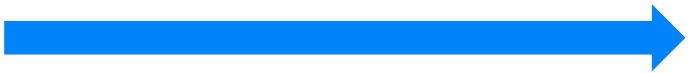


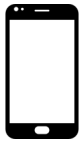
Software Vulnerability Study: Earbuds vs. Smartphone

- Limited calculation resources
- Tradeoff between quick response and security



Software Vulnerability Study: Earbuds vs. Smartphone





"Hey Siri"
"Hey Siri, send a message"
"Hey Siri, how's the weather today?"
"Hey Siri, set a timer for three minutes."
"Hey Siri, play some music."

Software Vulnerability Study: Earbuds vs. Smartphone

malicious voice commands with 47 different generated voices

"Hey, XX"

Pixel phone, iPhone

"Hey, XX"

Pixel buds, AirPods, Bose

Software Vulnerability Study: Earbuds vs. Smartphone

malicious voice commands with 47 different generated voices

"Hey, XX"

Pixel phone, iPhone

Rejected

0%

"Hey, XX"

"Calling mom"

26~100%

Pixel buds, AirPods, Bose

Outline

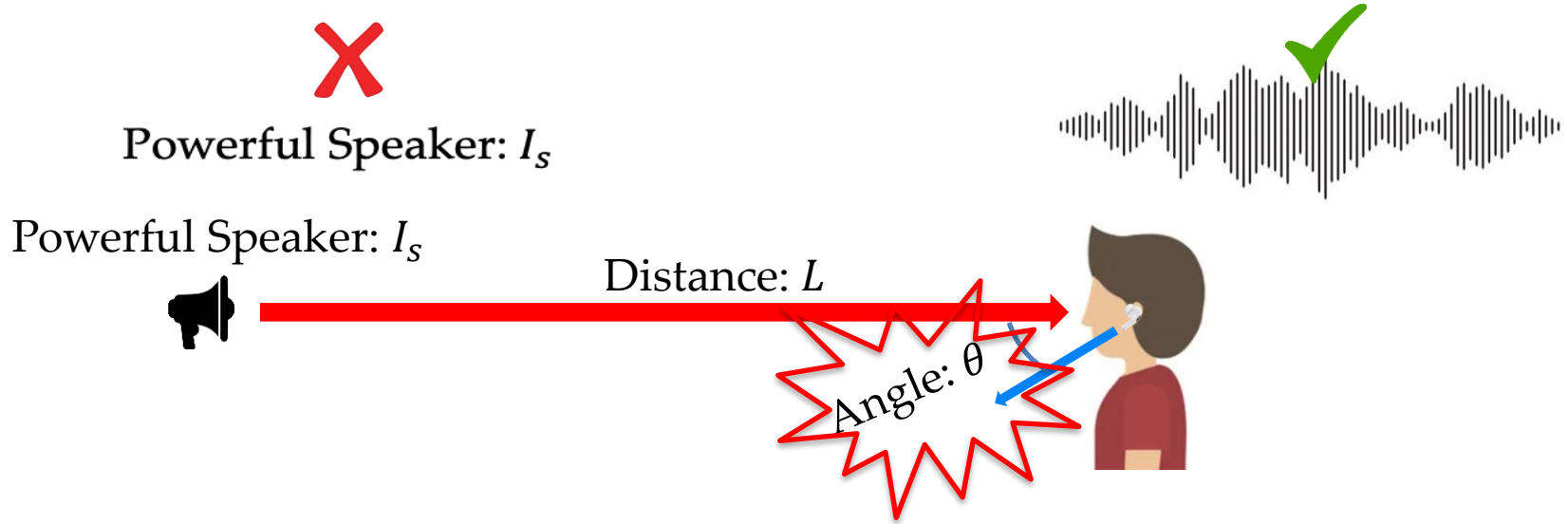
Practical Inaudible Attacks
(MobiSys'23)

Multi-Modal Authentication Framework
(MobiSys'21, Ubicomp'23, MobiSys'24, NDSS'25)

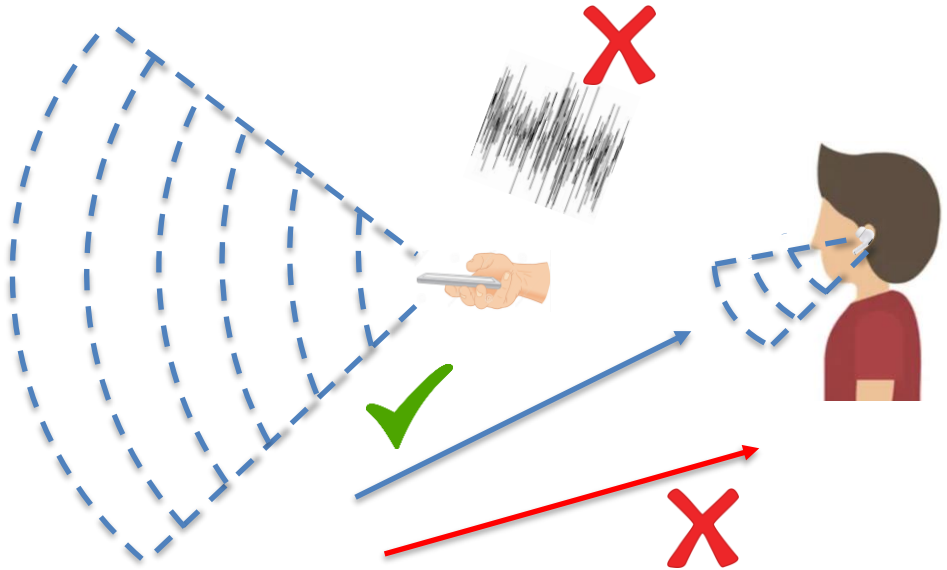
Can we attack earbuds practically?

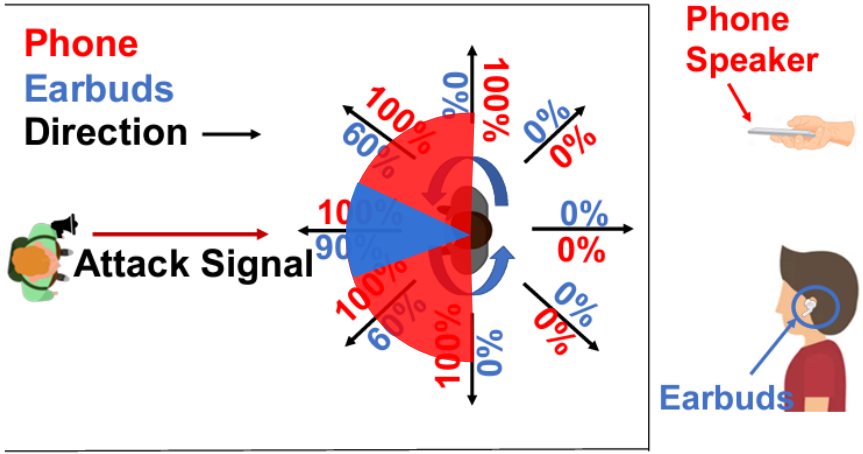
Challenge #1: Low Signal-to-noise Ratio (SNR)

Challenge #1: Low SNR



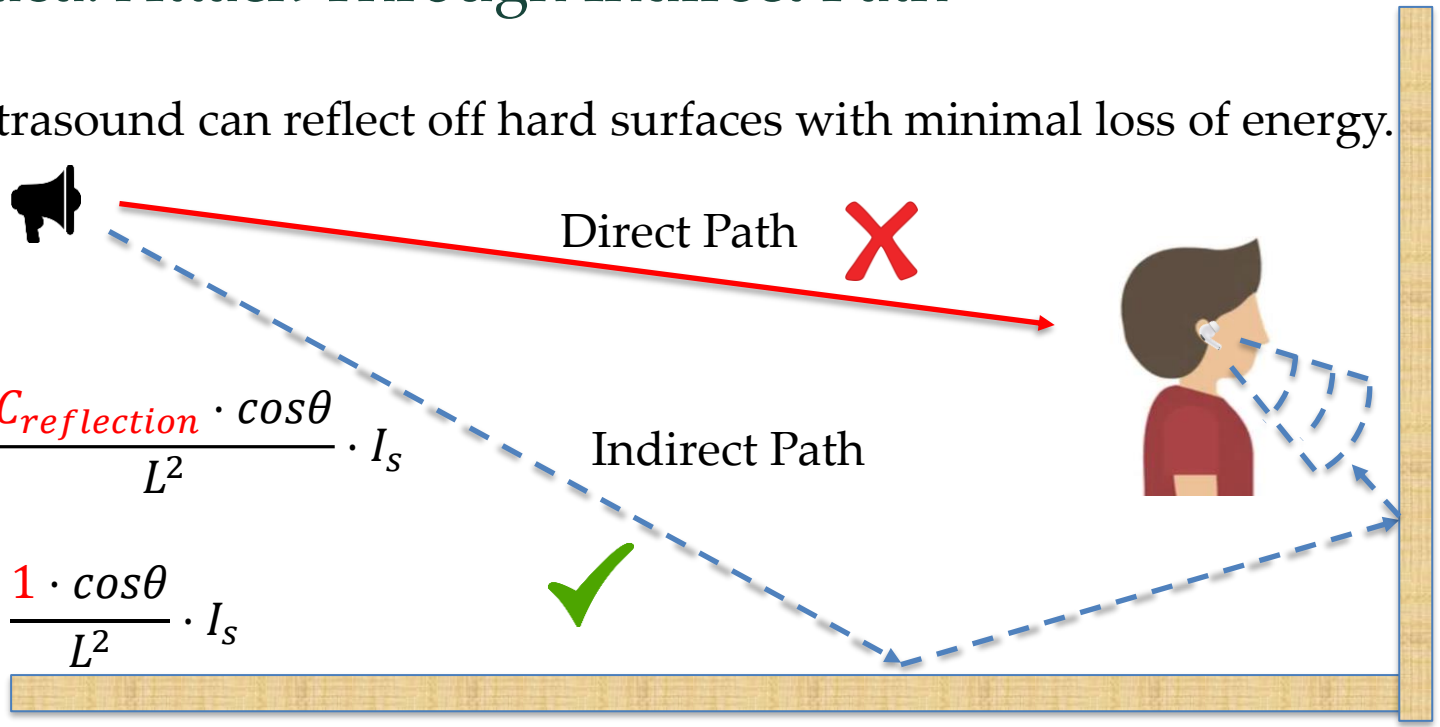
Challenge #1: Low SNR





Key Idea: Attack Through Indirect Path

Ultrasound can reflect off hard surfaces with minimal loss of energy.

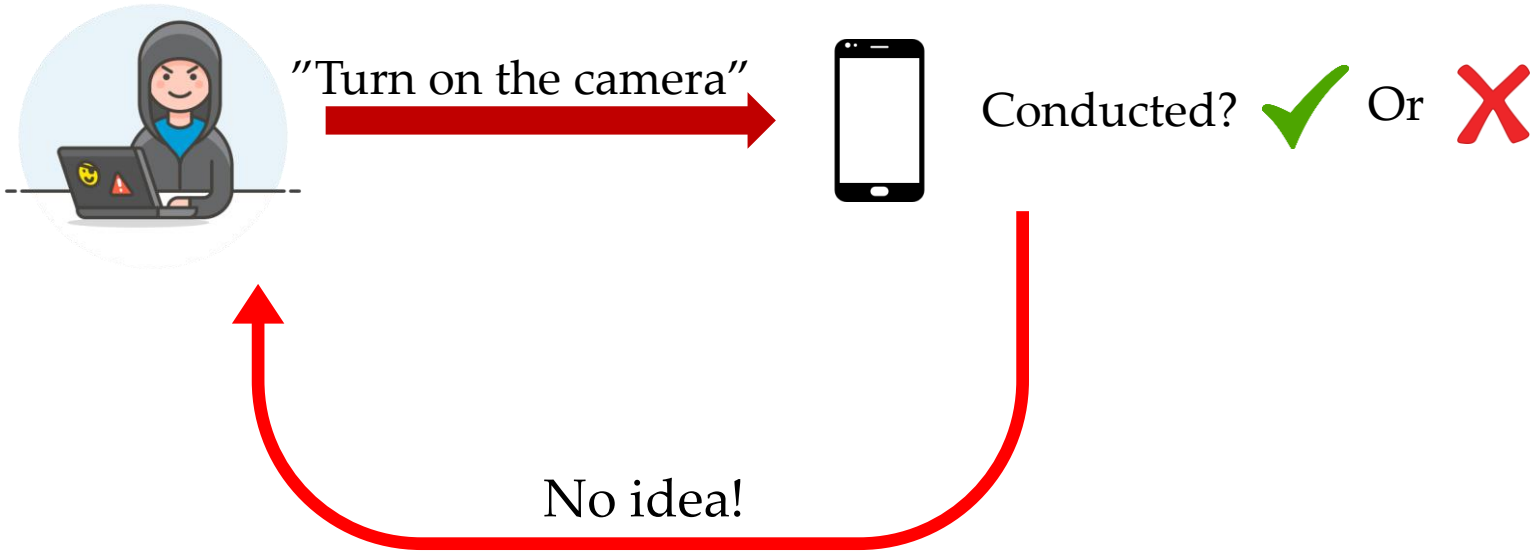


$$I_{received} = \frac{C_{reflection} \cdot cos\theta}{L^2} \cdot I_s$$

$$\approx$$

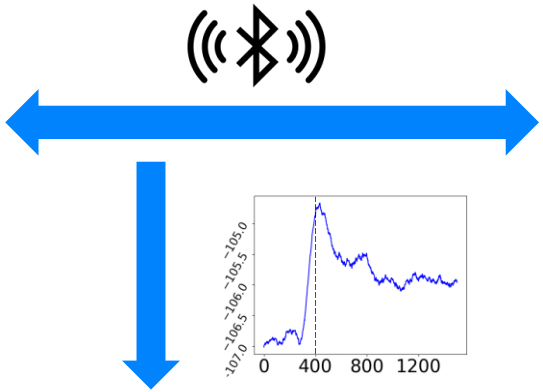
$$I_{received} = \frac{1 \cdot cos\theta}{L^2} \cdot I_s$$

Challenge #2: Unnoticeable Feedback



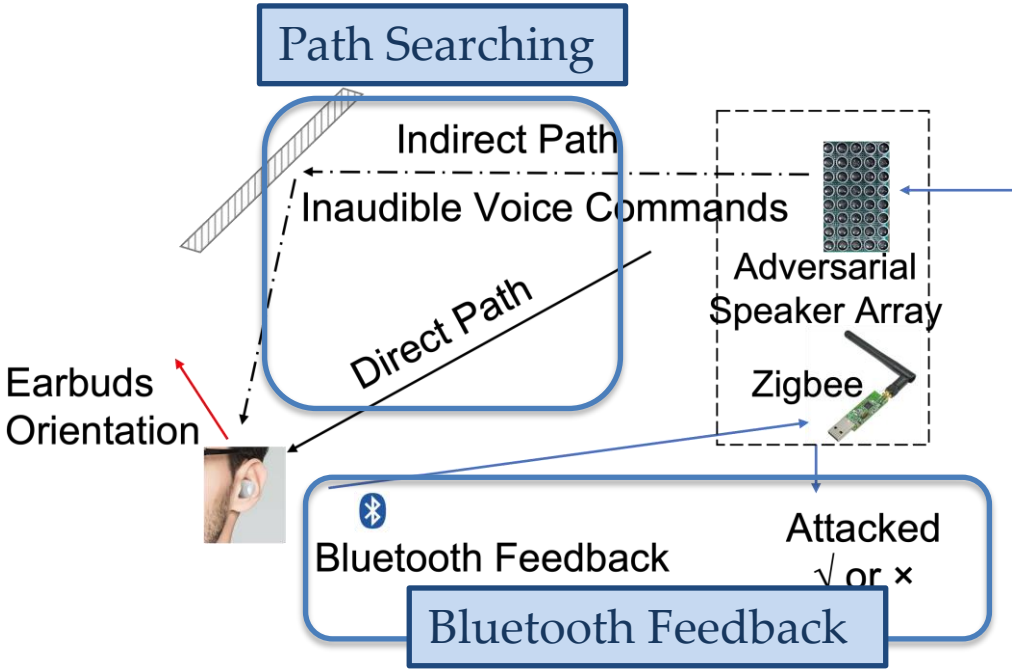
Key Idea：Utilize Bluetooth Side Channel

"Ok Google"



Conducted?

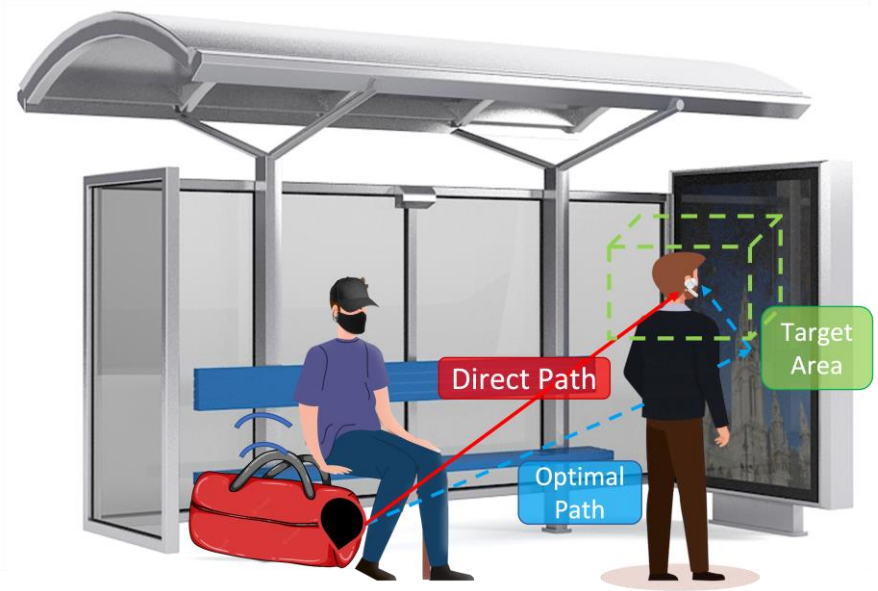
EchoAttack Overview



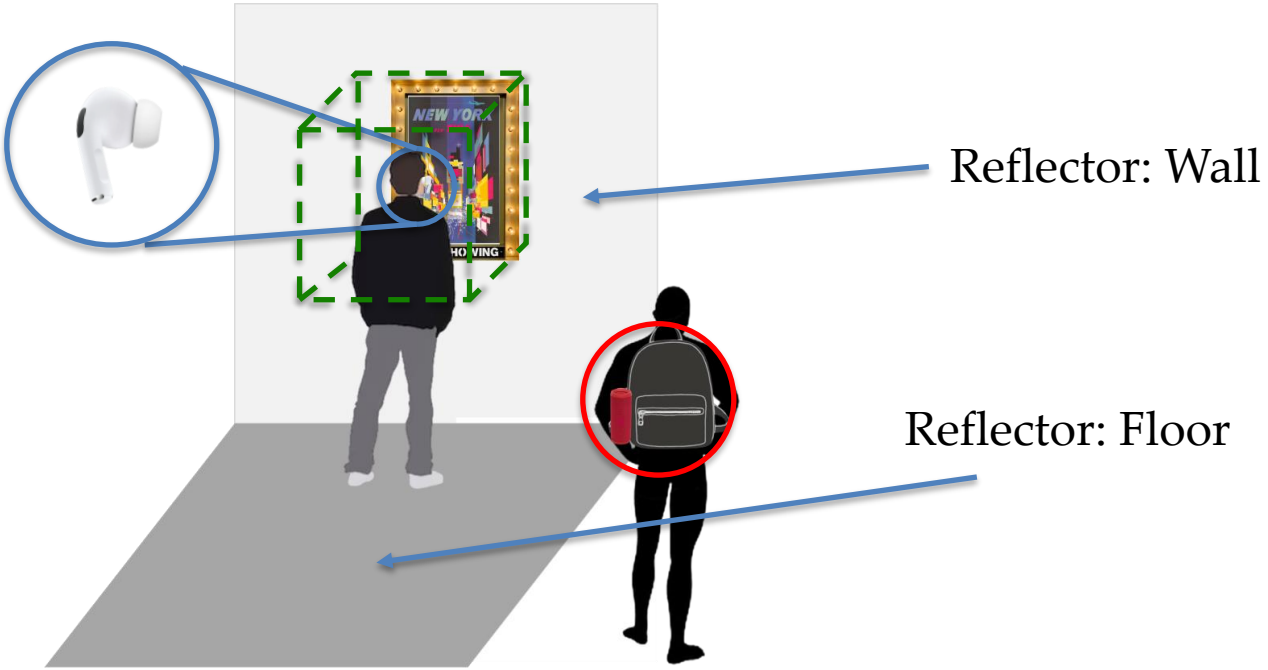
System Design

Path Searching

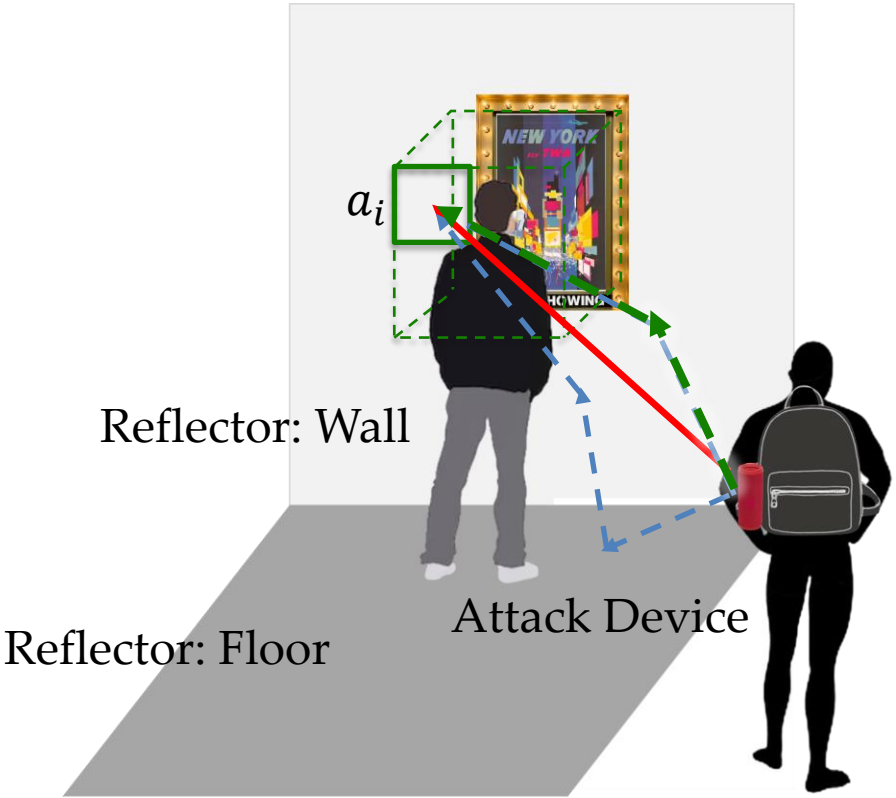
Bluetooth Feedback



Threat Model



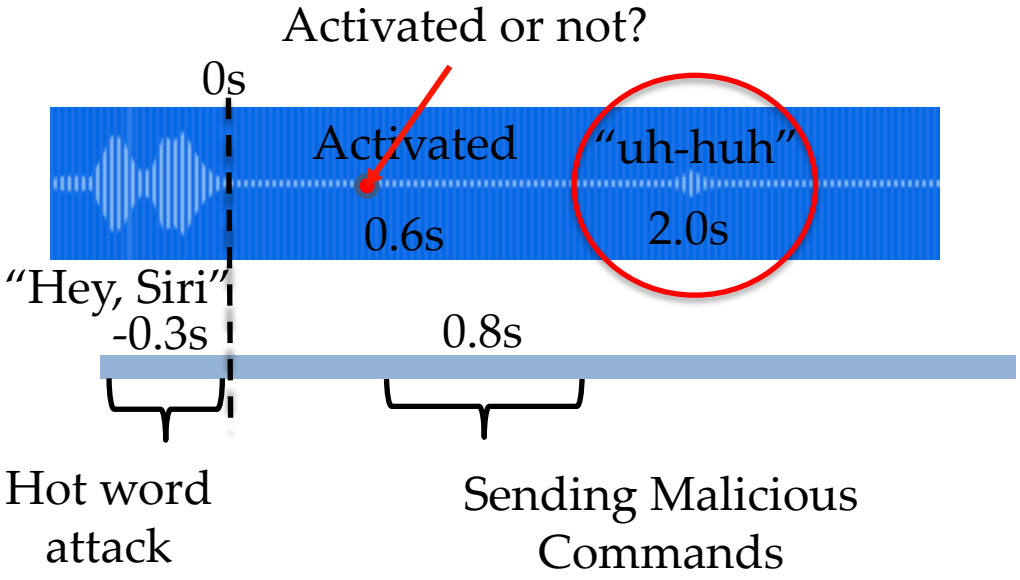
Path Searching



Select the path that keeps the most energy

$$Max(I_{received} = \frac{cos\theta}{L^2} \cdot I_s)$$

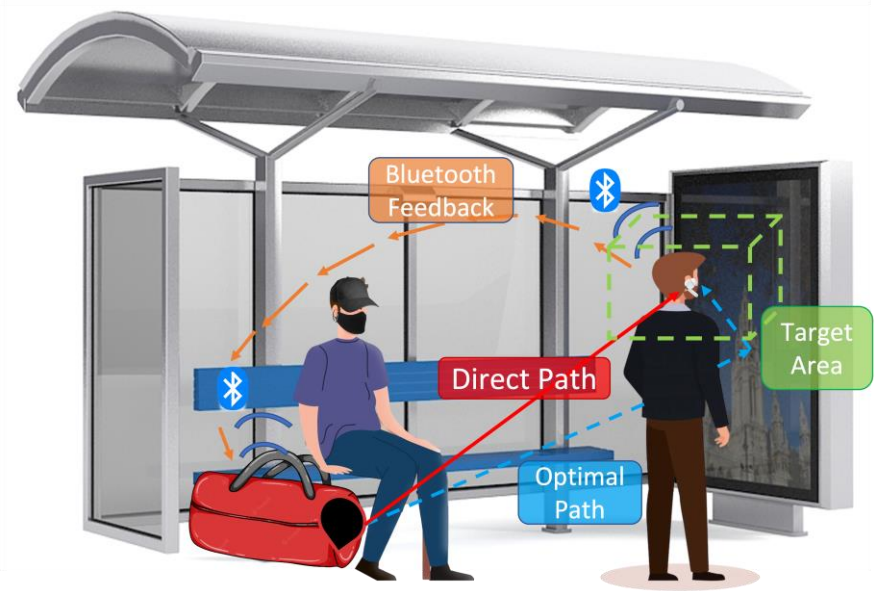
Launching Ultrasound Attacks



System Design

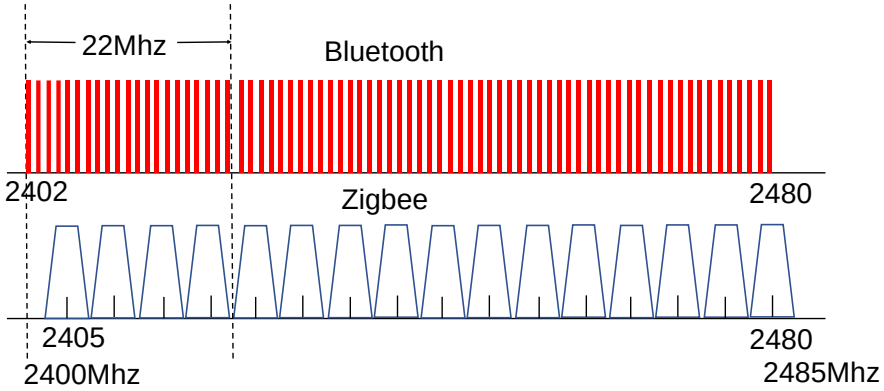
Path Searching

Bluetooth Feedback



Bluetooth Feedback

How to get energy change in Bluetooth Channel?

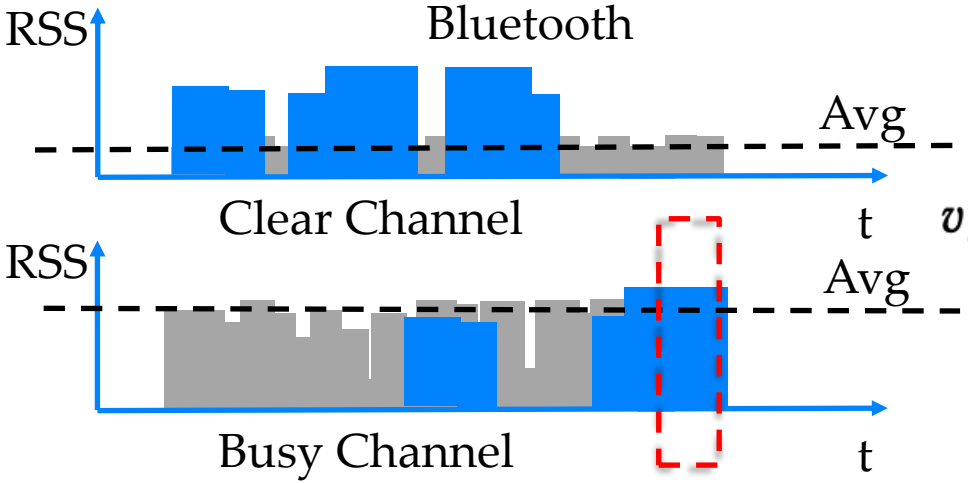


1 master+16 workers
Received signal strength(RSS)

Bluetooth Feedback

How to eliminate noise of Wi-Fi and Zigbee?

Bluetooth: Adaptive frequency hopping

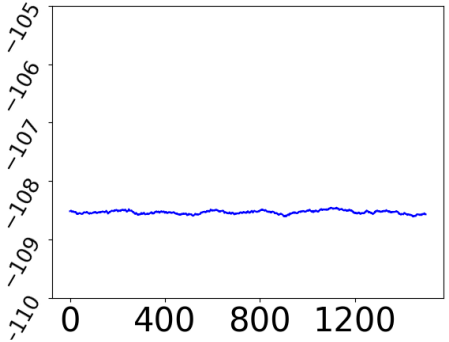


$$v_{ij} = \begin{cases} -110, & v_{ij} < average(V_{noise}[i]) \\ v_{ij}, & v_{ij} \geq average(V_{noise}[i]) \end{cases}$$

Bluetooth Feedback

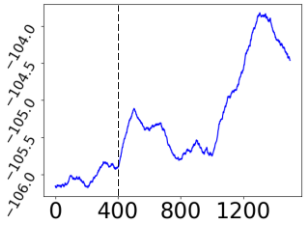
Siri, Alexa, Pixel, Bixby

Not activated: 100
Actived: 100

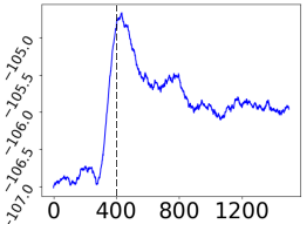


Bluetooth Feedback

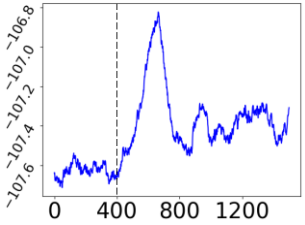
"Ok Google"



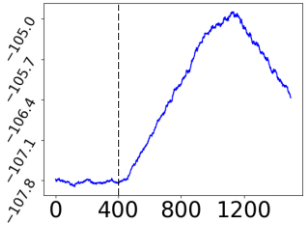
Siri



Pixel



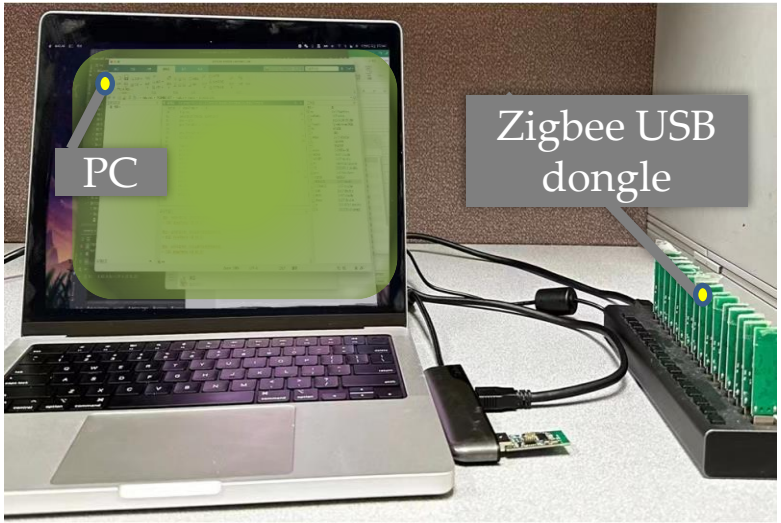
Alexa



Bixby

EchoAttack Implementation

EchoAttack Implementation



EchoAttack Evaluation

Earbuds: 

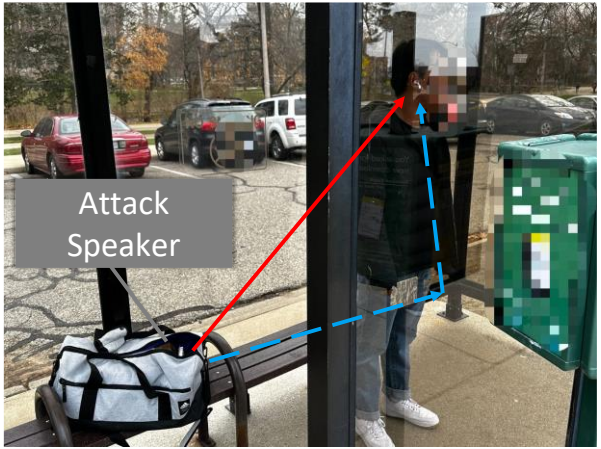
Operate system: 

Voice Assistant: 

Baseline: Using only direct path (**Direct Attack**)

EchoAttack Evaluation: Three Real-world Scenarios

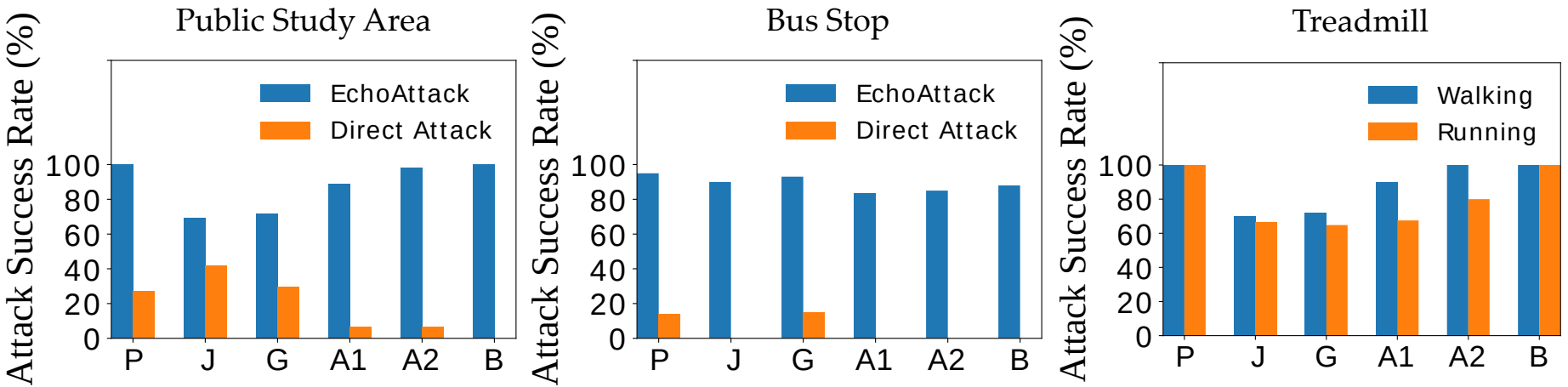
Public study area



Bus stop

Gym

EchoAttack Evaluation: Overall Performance

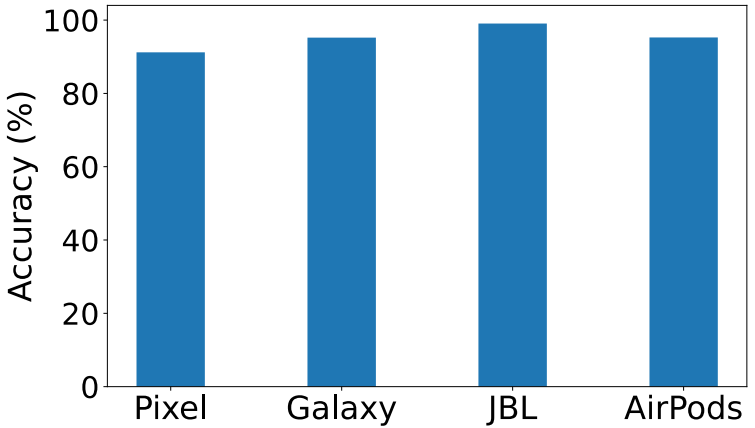


EchoAttack: 85% Direct Attack: 15%

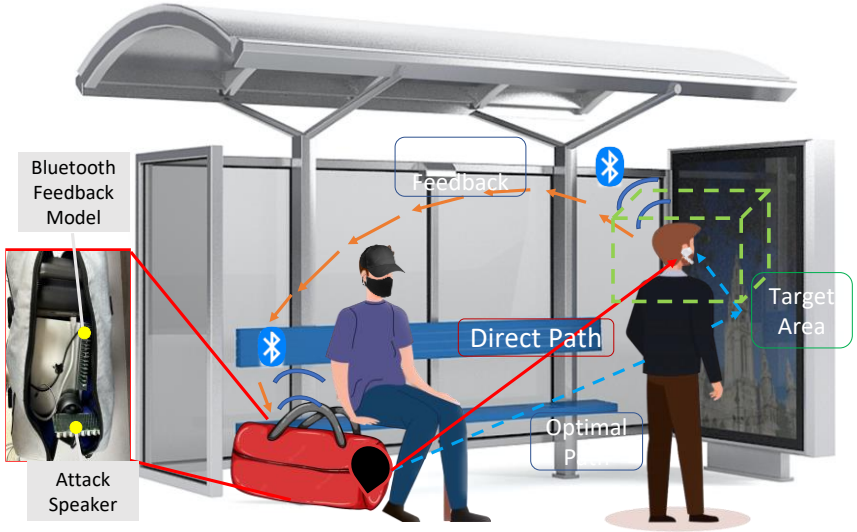
EchoAttack Evaluation: Bluetooth Feedback Module

Activated : Non-Activated = 100 : 100

Average Accuracy: 90%+



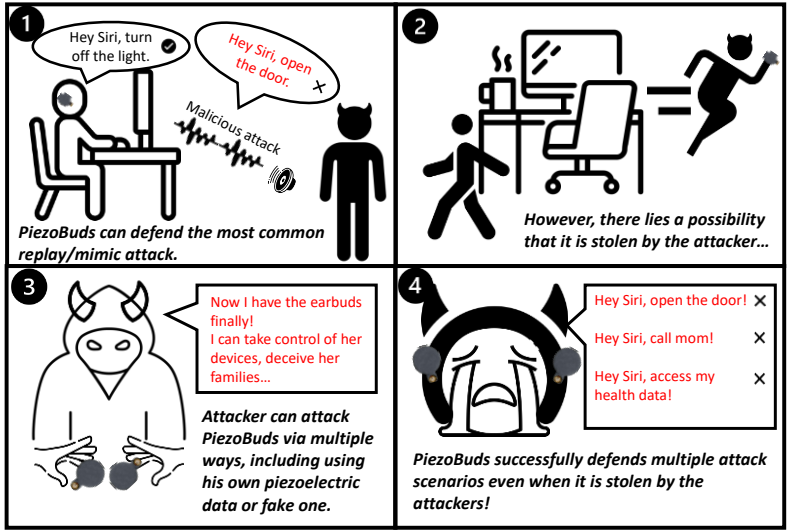
Takeaways



- We reveal the ultrasound attack threats to the voice assistants on earbuds.
- We design EchoAttack to conduct ultrasound attack over earbuds.
- EchoAttack achieves a high attack success rate of about 85% on average.

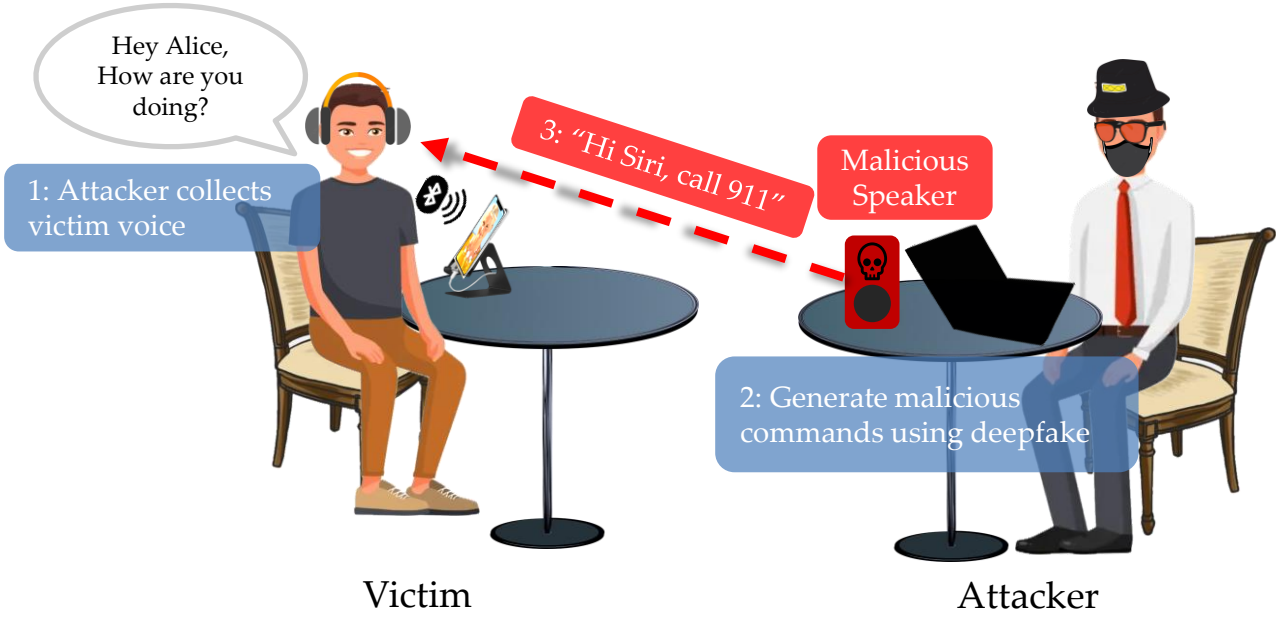
Outline

Practical Inaudible Attacks
(MobiSys'23)

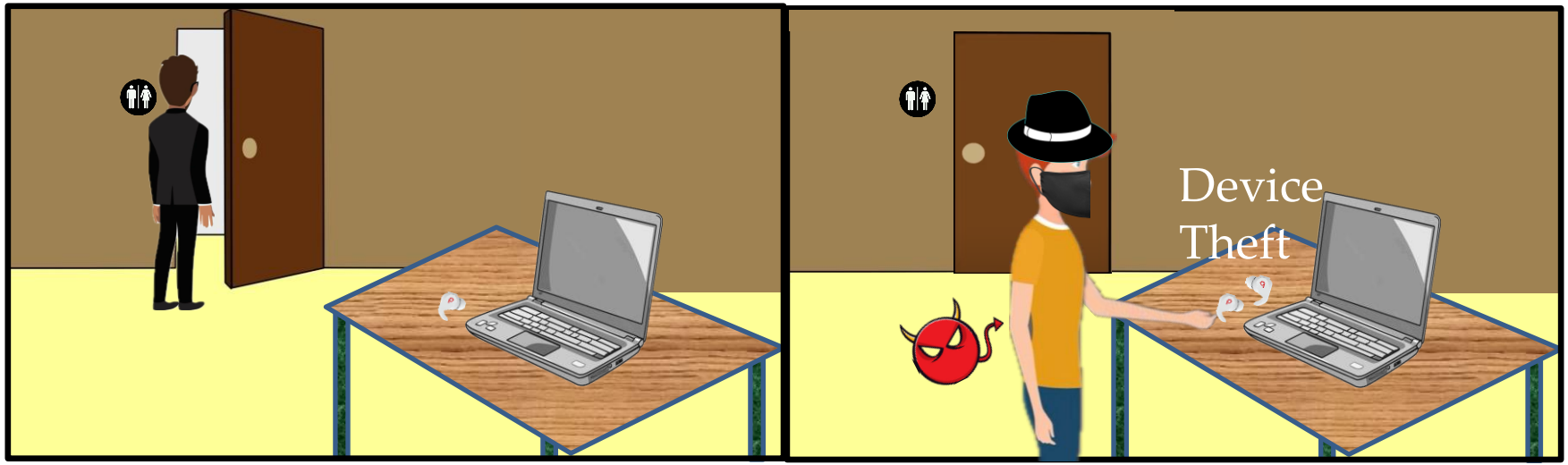


Multi-Modal Authentication Framework
(MobiSys'21, Ubicomp'23, MobiSys'24, NDSS'25)

Audio Deepfake Attacks



Access The Victim Devices



Threat Model

- Synthesis Attack.
- Replay Attack.
- Audio Deepfake Attack.
- Hybrid Attack.

Existing Multi-modal Earable Authentication

Accuracy ~ 95% 😕

Practical Considerations 😕

Computation Overhead 😕

Audio + In-ear sound [1]

Audio + IMU [2]

Audio + PPG [3]

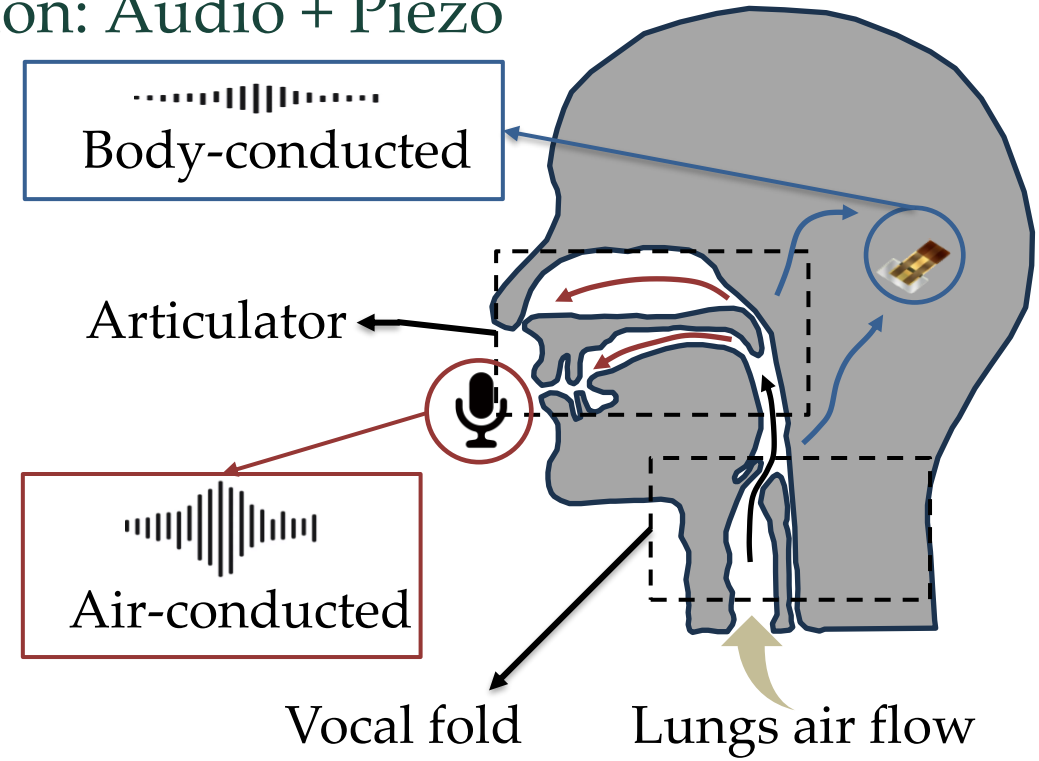
[1] Z. Wang, S. Tan, L. Zhang, Y. Ren, Z. Wang, and J. Yang, "Eardynamic: An ear canal deformation based continuous user authentication using in-ear wearables," Proceedings of Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 5, no. 1, 2021.
[2] J. Liu, W. Song, L. Shen, J. Han, and K. Ren, "Secure user verification and continuous authentication via earphone imu," IEEE Transactions on Mobile Computing, vol. 22, no. 11, pp. 6755–6769, 2023.
[3] S. Choi, J. Yim, Y. Jin, Y. Gao, J. Li, and Z. Jin, "Earppg: Securing your identity with your ears," in Proceedings of ACM IUI, 2023.

Is there an accurate, practical, and lightweight anti-attack authentication for earbuds?

Our Solution: Audio + Piezo

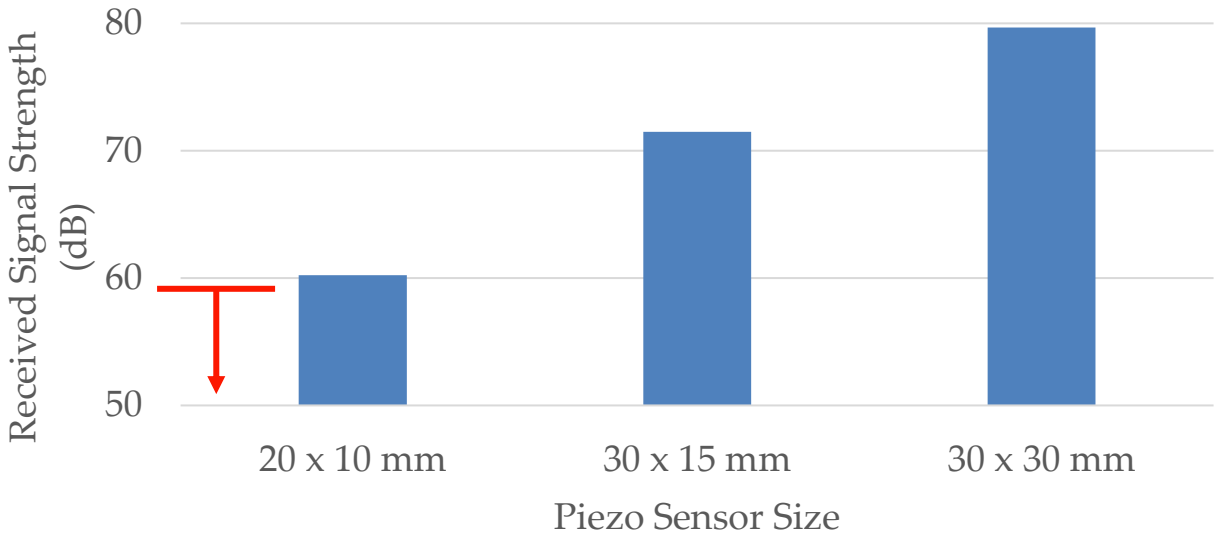
Piezoelectric sensor

Our Solution: Audio + Piezo

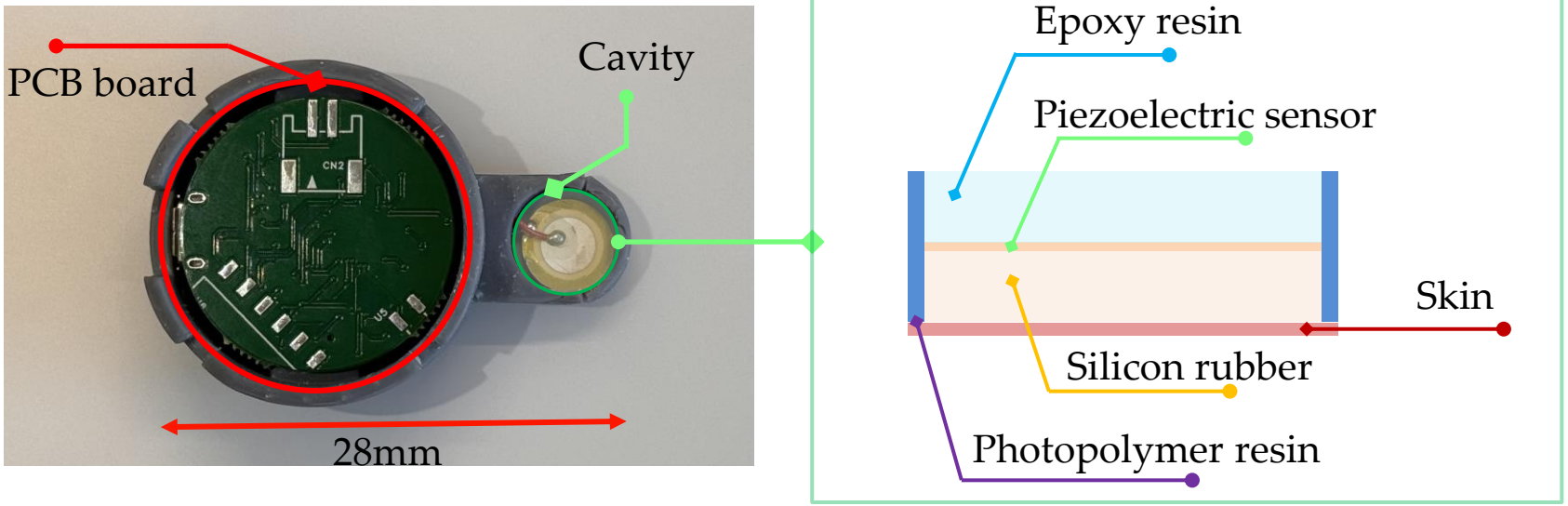


Challenge #1: Low Signal-to-noise Ratio (SNR)

- Sensor size ≈ sensor performance



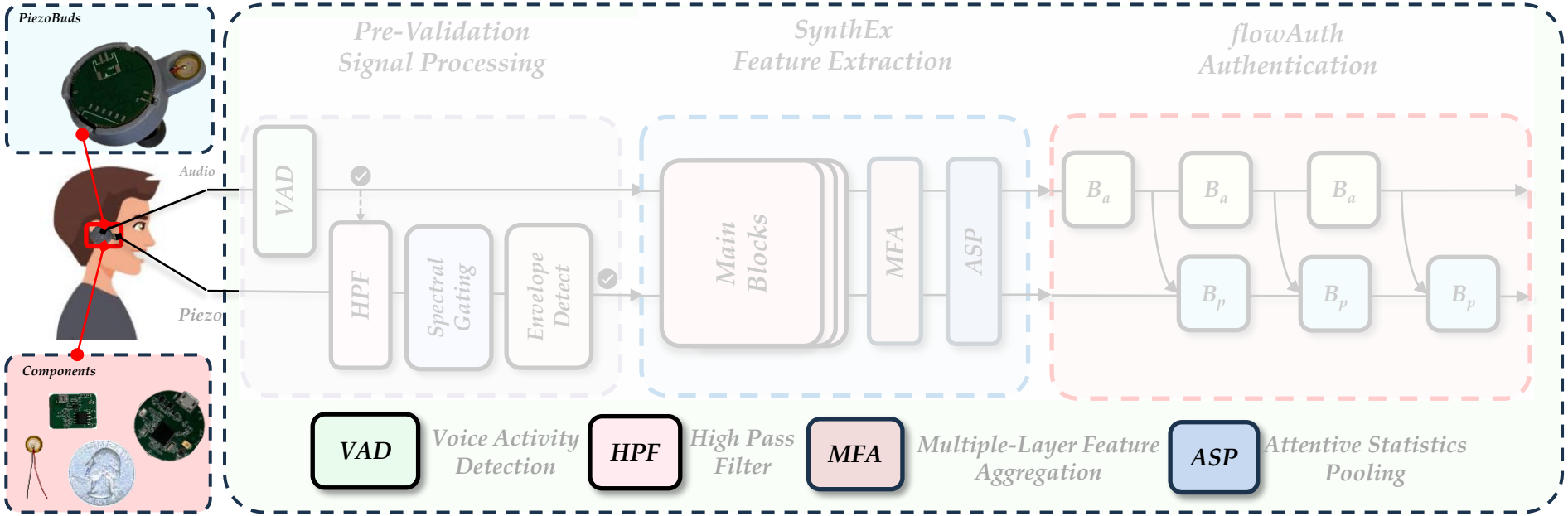
Our Solution: Multi-layer Piezo Sensing Structure



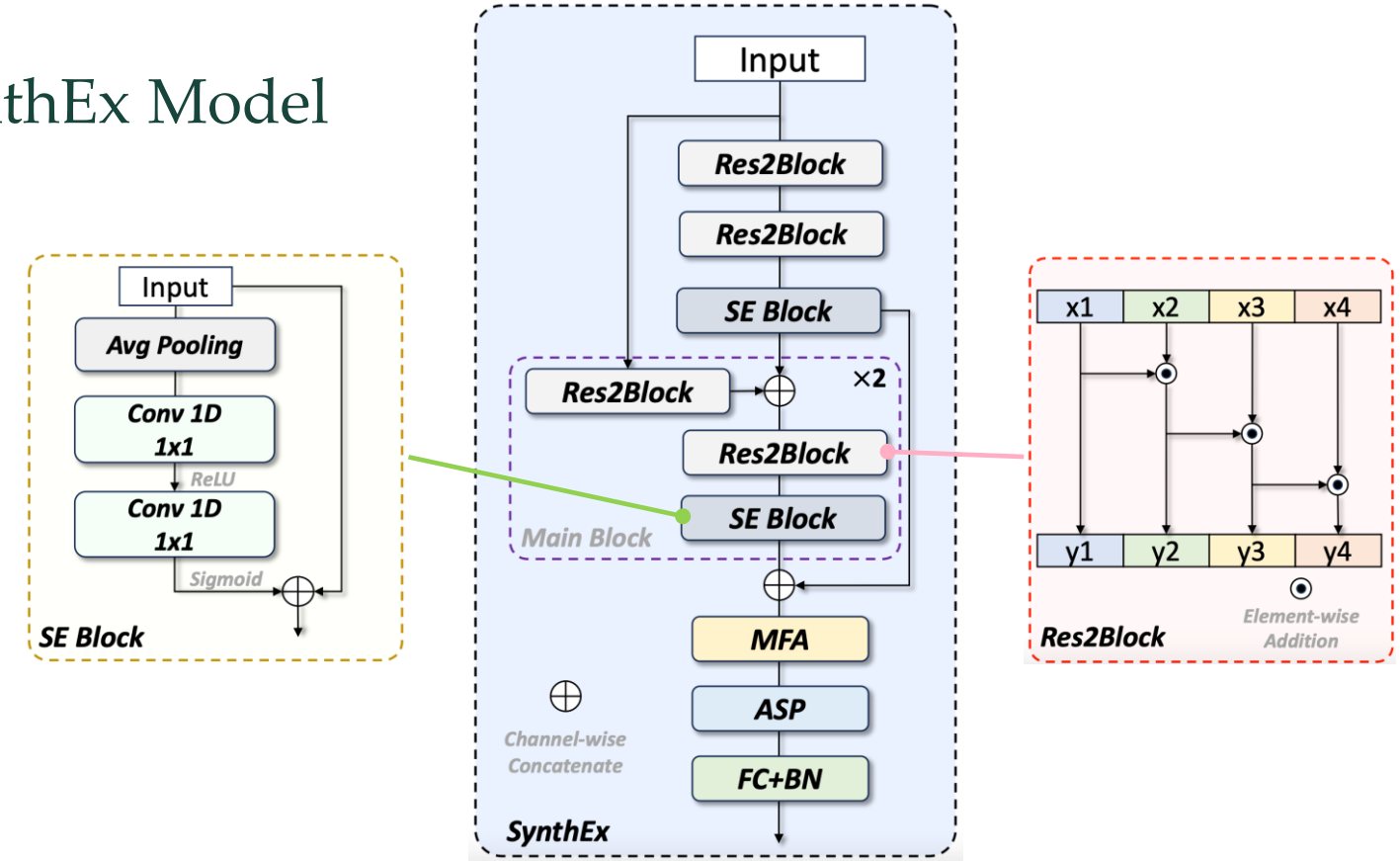
Challenge #2: How to fuse two sensing modalities for authentication?

- Heterogeneous sensor data
- Computation overhead

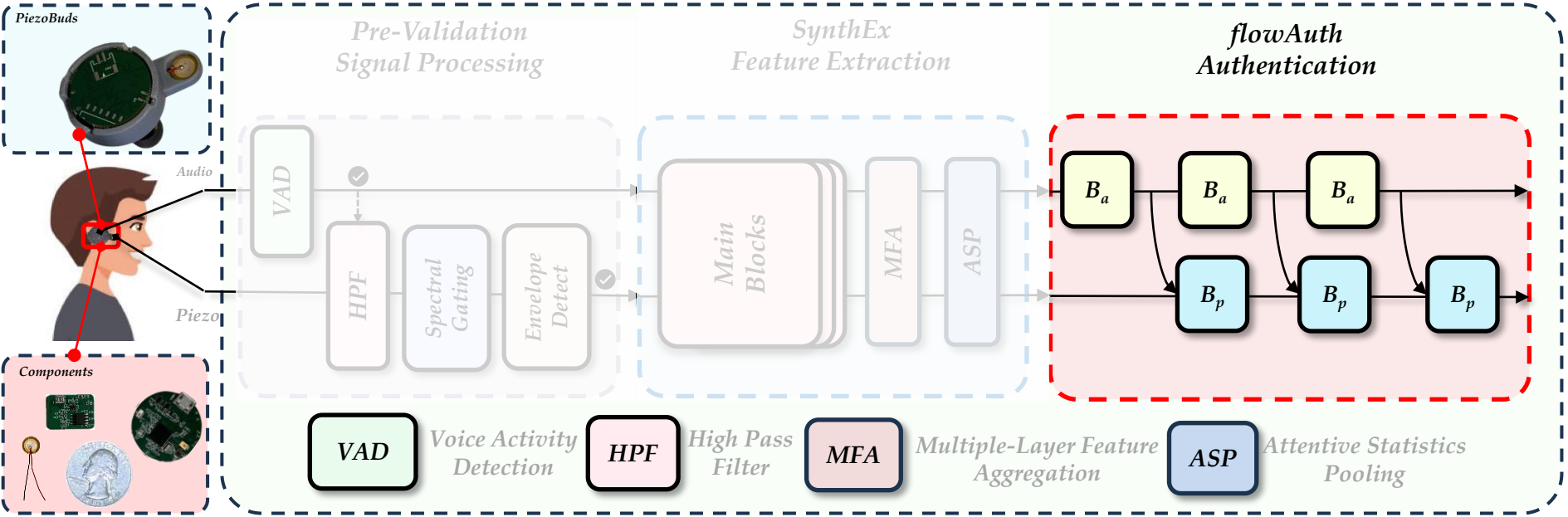
Our Solution: FusionSecNet



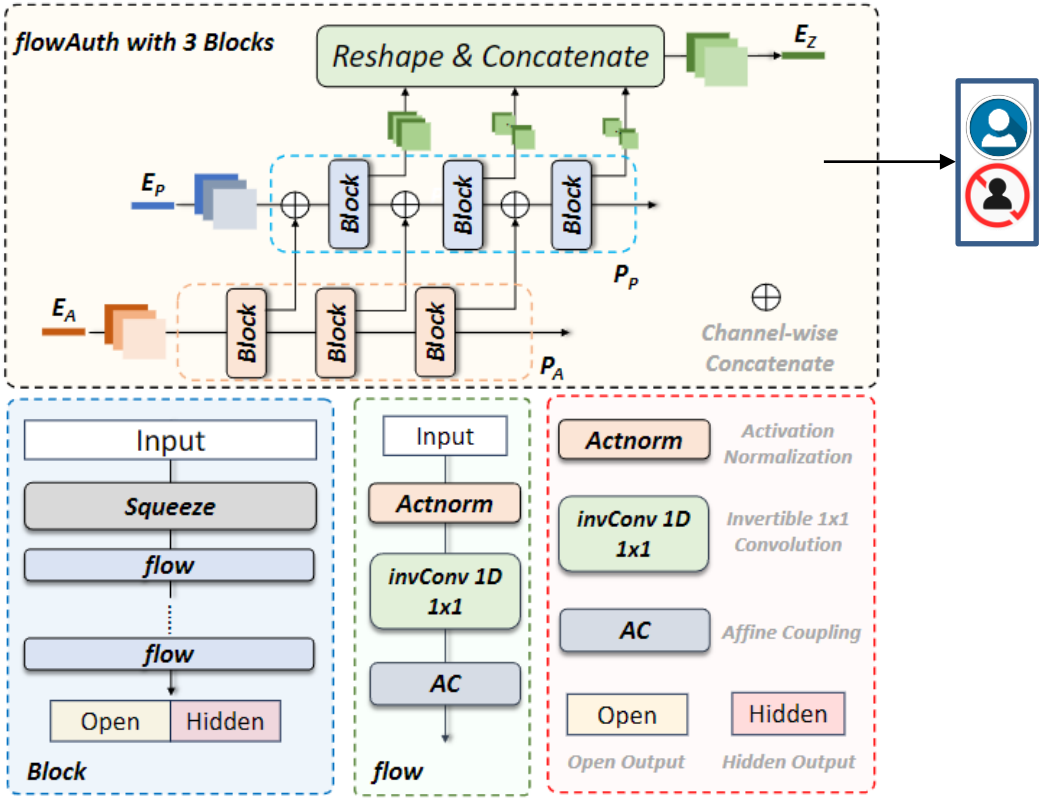
SynthEx Model



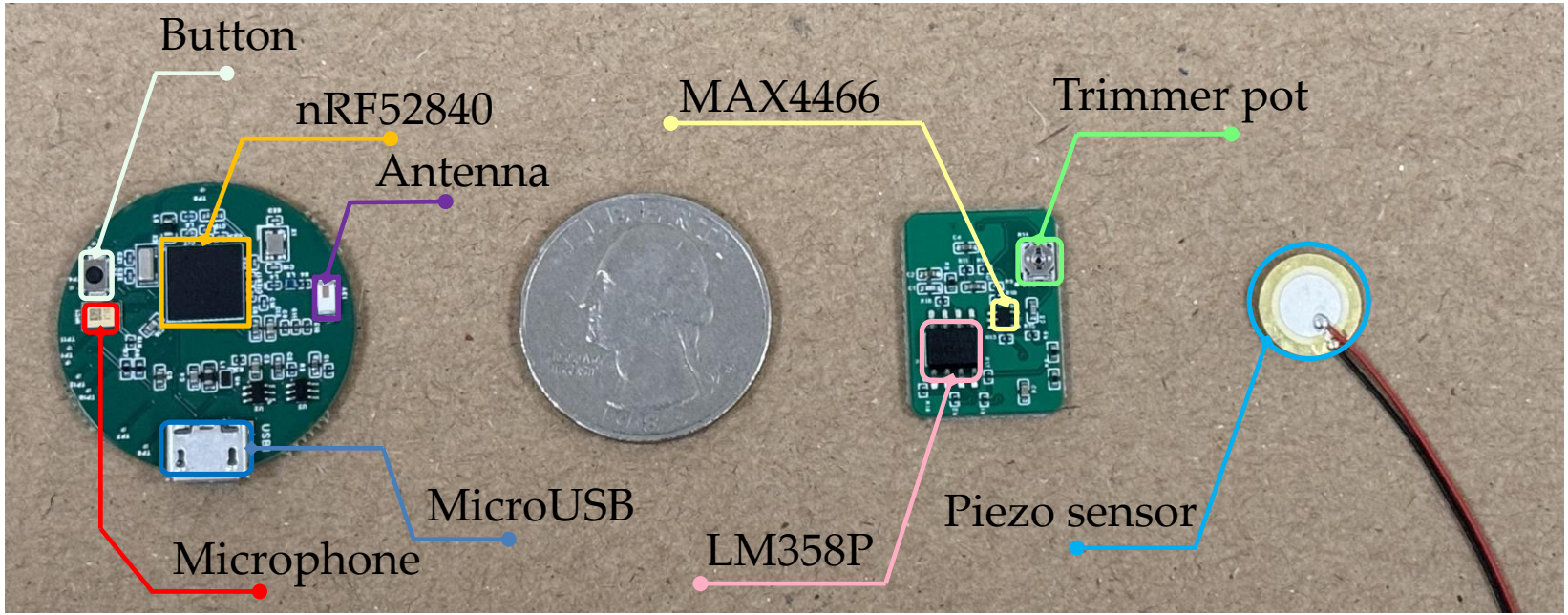
Our Solution: FusionSecNet



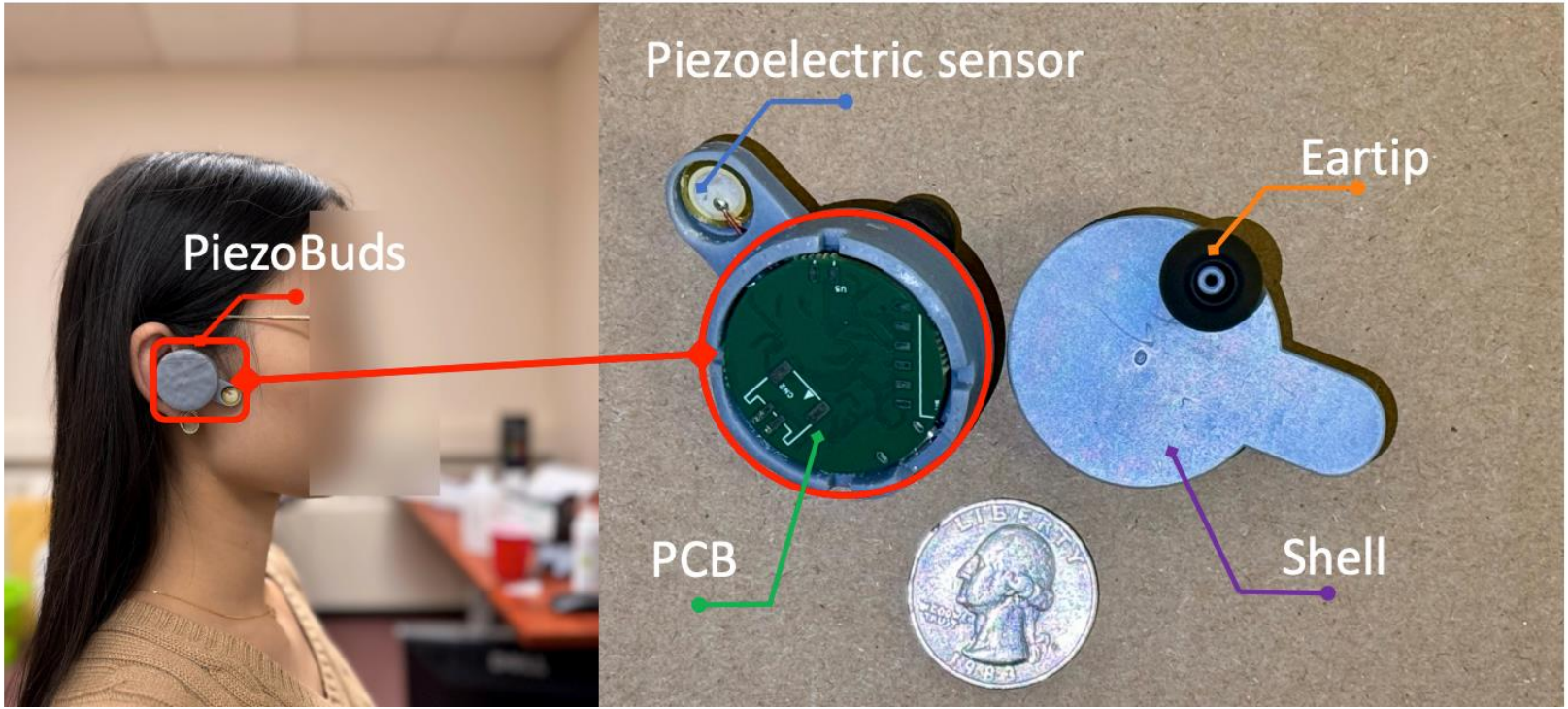
FlowAuth Authentication



PiezoBuds Implementation



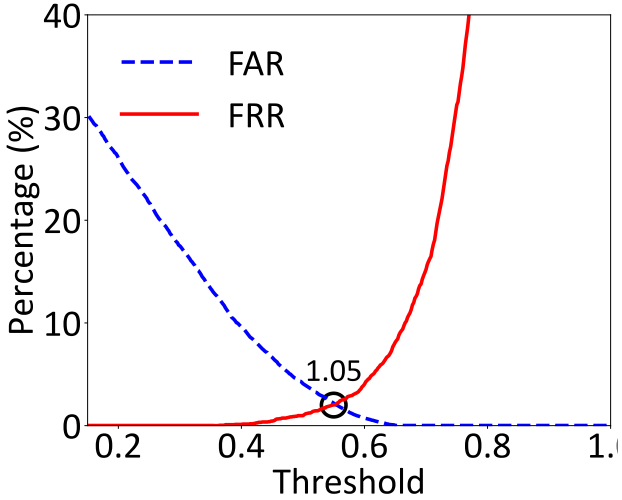
PiezoBuds Implementation

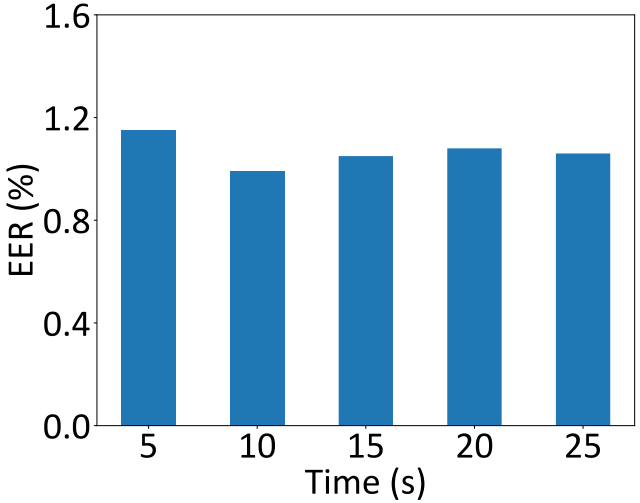


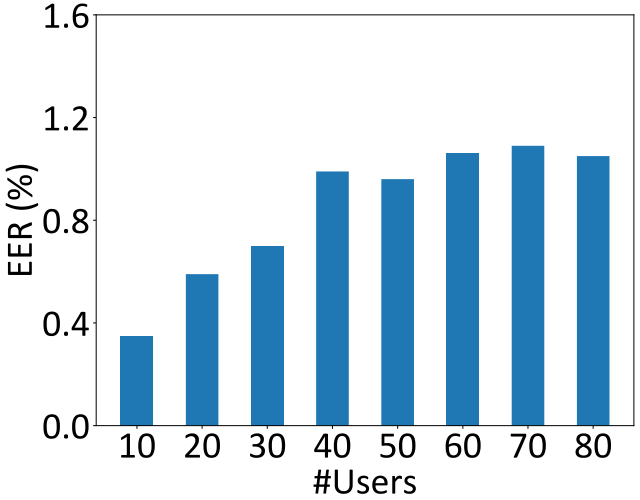
Experimental Settings

- 85 participants (64 M, 21 F, 45 native speakers)
- Reading materials
  - 3 types (content is random): a speech script, a fairy tale, and scientific-educational content
  - 1000 word per text
- Evaluation Metrics
  - Equal Error Rate (EER)

Authentication Performance







Anti-attack Performance

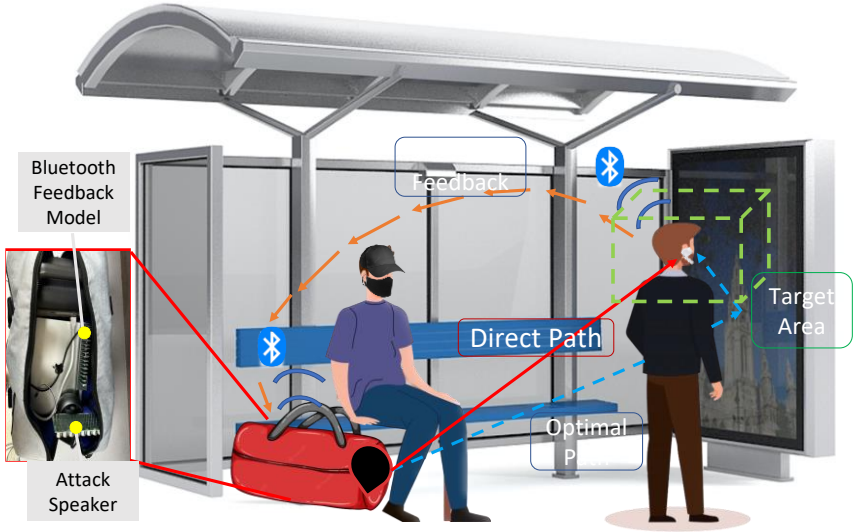
<table>
<tr><th colspan="3">Attack Scenarios</th><th>Defense Success Rate</th></tr>
<tr><td rowspan="4">Scenario 1</td><td colspan="2">Replay only</td><td>100%</td></tr>
<tr><td rowspan="3">Mimic</td><td>ResembleAI [1]</td><td rowspan="3">100%</td></tr>
<tr><td>PlayHT [2]</td></tr>
<tr><td>Vall-E [3]</td></tr>
<tr><td colspan="3">Scenario 2: Replay + Attacker’s Piezo</td><td rowspan="3">100%</td></tr>
<tr><td colspan="3">Scenario 3: Replay + Skull Generated Piezo</td></tr>
<tr><td colspan="3">Scenario 4: Replay + Generative AI Piezo</td></tr>
</table>

[1] Resemble, “Resemble.ai,” https://www.resemble.ai/, Retrieved by Apr 17 2024.
[2] play.hy, “Playht,” https://play.ht/, Retrieved by Apr 17 2024
[3] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, “Neural codec language models are zero-shot text to speech synthesizers,” 2023.

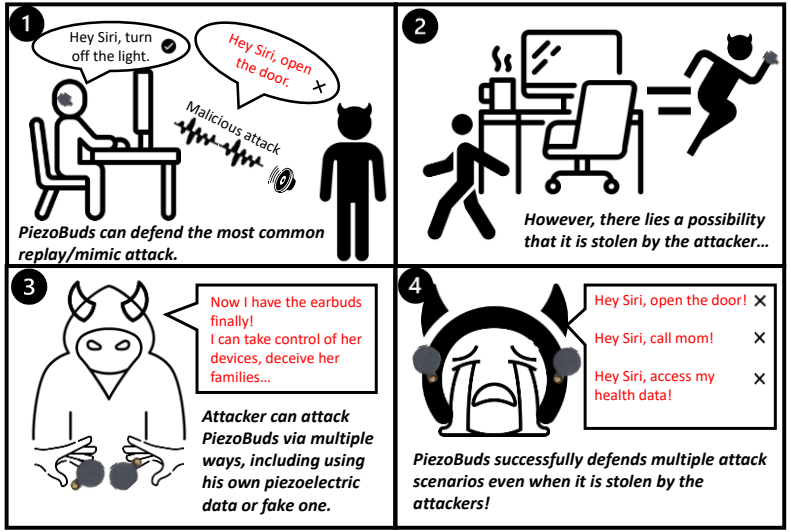
Comparison With The State-of-the-art Methods

	Modalit(ies)	Devices	# subjects	ACC (%) / EER (%)	Enrollment length (s)	Latency (s)	Defense Task	FAR (%)
[54]	Piezo + Audio	Headset	-	97 / -	N.A.	0.8 – 1.2	Live Check	2
[37]	Piezo + Audio	Wearable	29	87.5 – 96.1 / -	107	2.17 – 4.53	Authentication	3.1 – 3.6
[14]	Piezo + Audio	Wearable	18	95 / -	N.A.	0.3 – 0.83	Live Check	0
[16]	In/out ear sound	Earable	23	- / < 4	75	0.389 – 0.484	Verification	0
[17]	Ear canal	Earable	20	95.16 / -	120	-	Verification	0.18 – 0.22
[62]	Ear canal	Earable	24	97.38 / -	-	-	Authentication	5.3
[35]	IMU	Earable	34	- / 1.28	-	2	Verification	0
[5]	PPG + Audio	Earable	25	94.84 / -	-	-	Authentication	-
PiezoBuds	Piezo + Audio	Earable	**85**	**99.21 / 1.05**	**15**	**0.04 – 0.219**	Authentication	**0**

Takeaway



Practical Inaudible Attacks
(MobiSys'23)



Multi-Modal Authentication Framework
(MobiSys'21, Ubicomp'23, MobiSys'24, NDSS'25)

Collaborators and Research Teams