Speak Up, I'm Listening: Extracting Speech from Zero-Permission VR Sensors

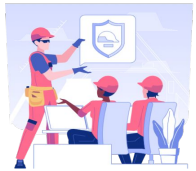
Derin Cayir[1], Reham Mohamed Aburas[2] , Riccardo Lazzeretti[3], Marco Angelini[4], Abbas Acar[1], Mauro Conti[5], Z. Berkay Celik[6], and Selcuk Uluagac[1]

[1]Florida International University, [2]American University of Sharjah, [3]Sapienza University of Rome, [4]Link Campus University of Rome, [5]University of Padua, [6]Purdue University

Virtual Reality (VR)

VR devices and their sensors are advancing, and their use cases are expanding to more sensitive scenarios:

Business meetings



Manufacturing

Gaming

Military trainings

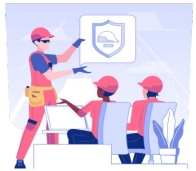
Medical field

Education

Virtual Reality (VR)

VR devices and their sensors are advancing, and their use cases are expanding to more sensitive scenarios:

Business meetings



Manufacturing

Gaming

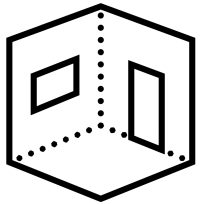
Military trainings

Medical field

Education

This is attributed to the sensors that perform:

Hand tracking



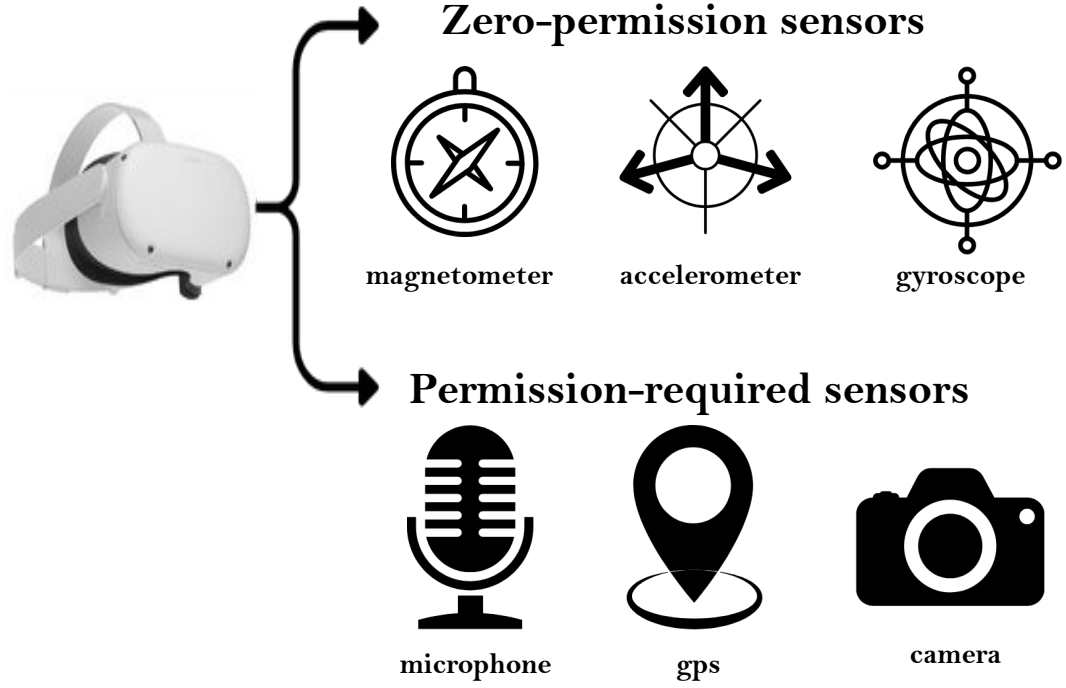
Environment mapping

User movement understanding

Eye tracking

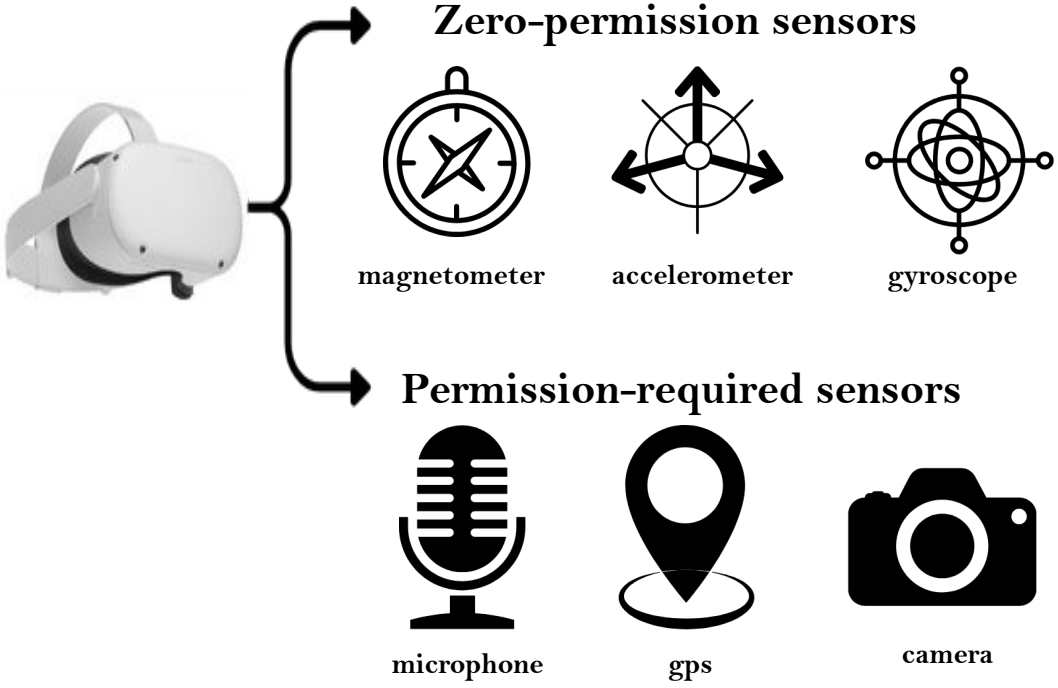
Zero-permission sensors

Sensors in VR can be categorized into two [1-4]:



Zero-permission sensors

Sensors in VR can be categorized into two [1-4]:

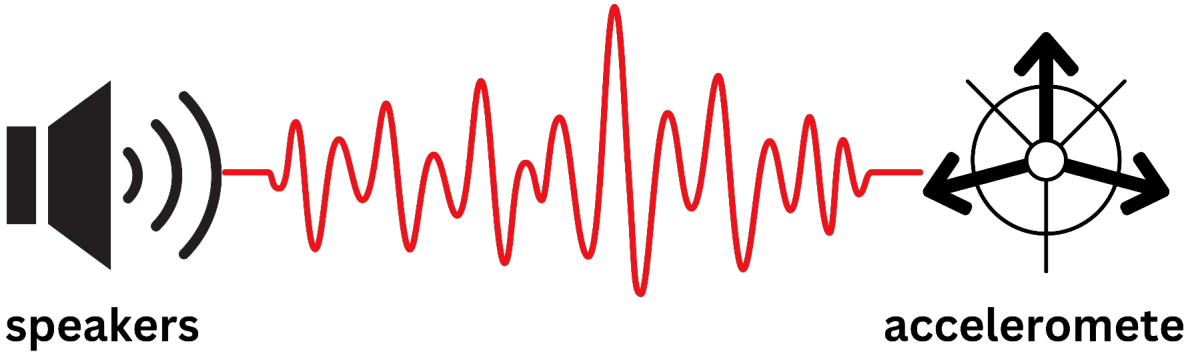


Accelerometers, while designed to serve for benign purposes (calculating the positional changes), still collect sensitive data [5-9].

Acoustic signals' effects on sensors

- Sound travels through vibrations and affect the accelerometer values [10].

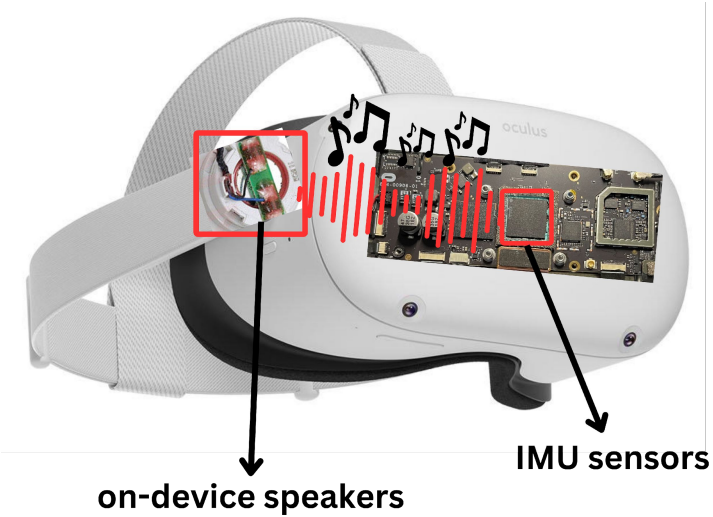
- These accelerometer values can be analyzed to recover the spoken content of the user.



Acoustic signals' effects on VR sensors

Problem:

- VR devices have **stronger speakers**, and **more capable accelerometer sensors** (reaching up to 1000 Hz whereas smartphones have 250Hz).
- VR captures **whole body movements.**
- **No defense** is explored other than reducing the sensor sampling rates.

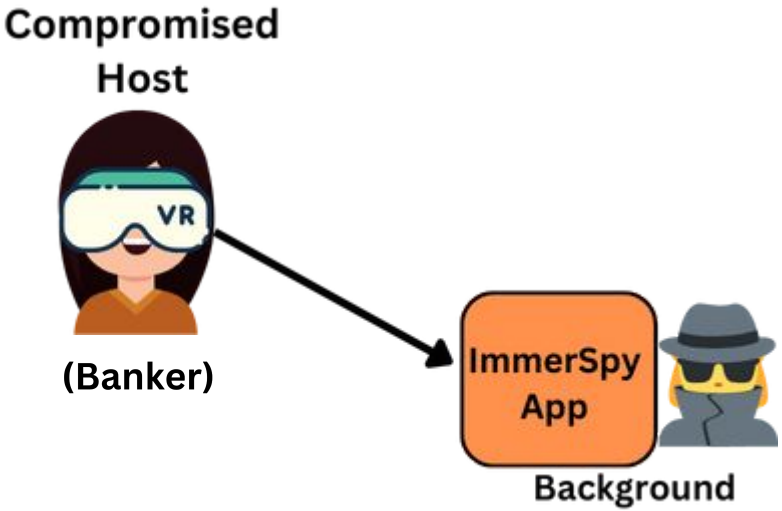


Threat model

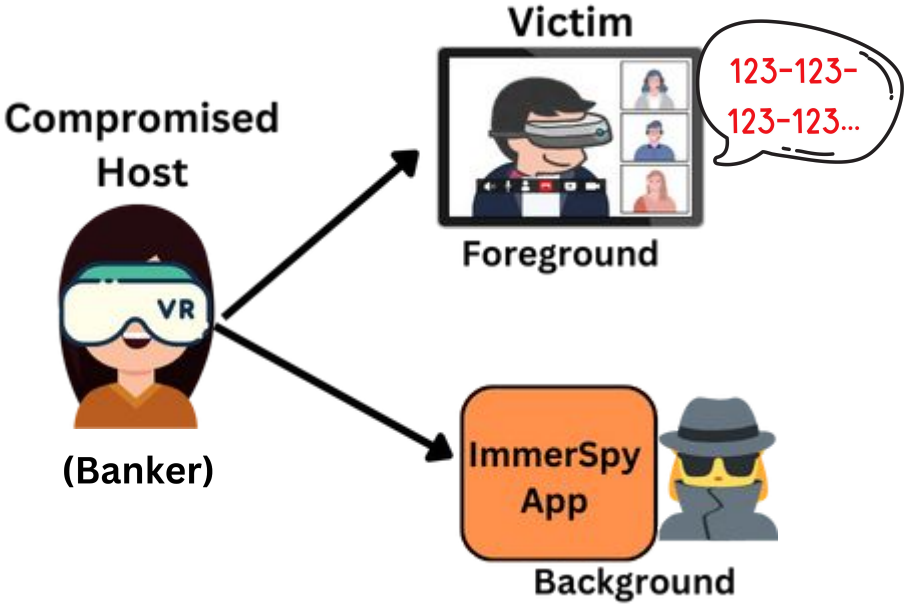
Compromised Host

(Banker)

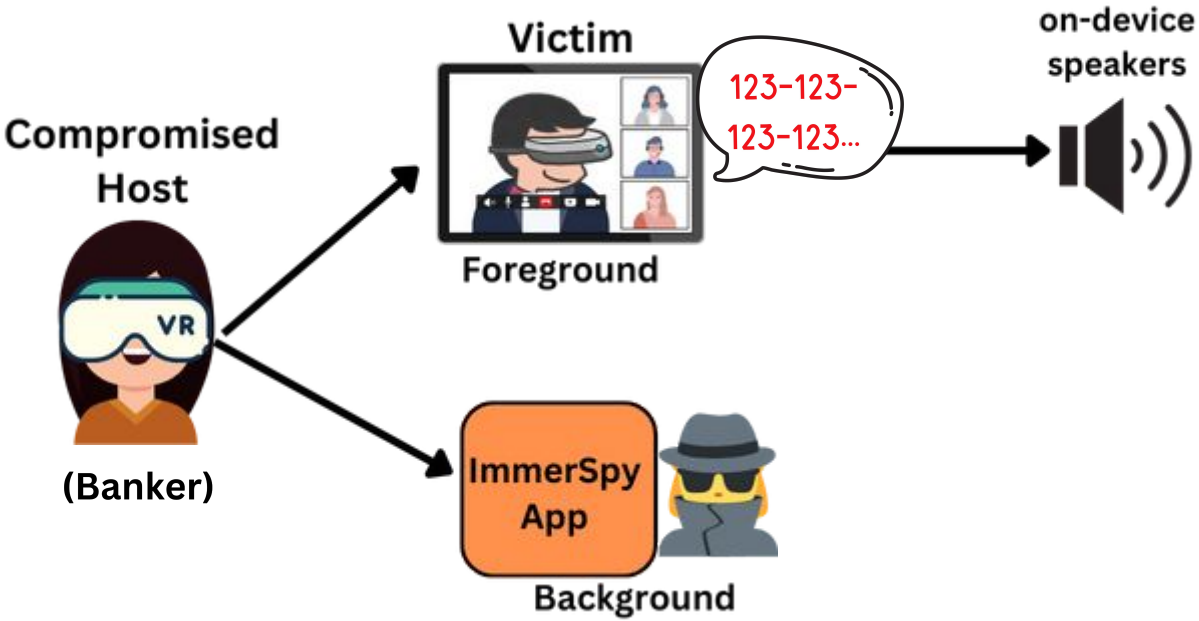
Threat model



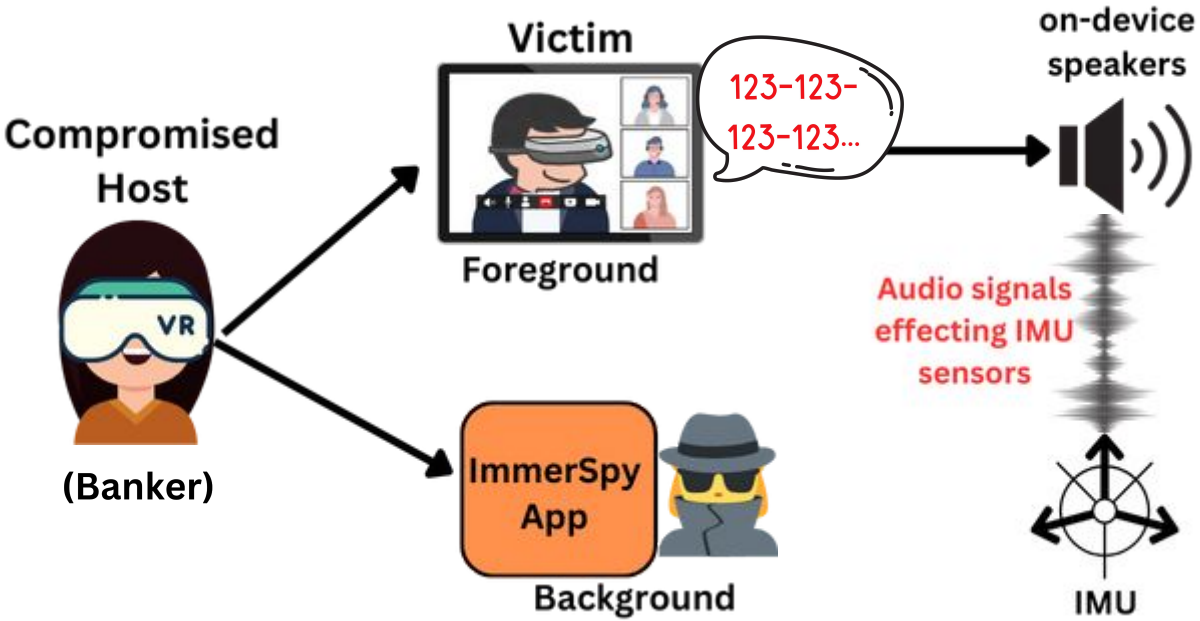
Threat model



Threat model

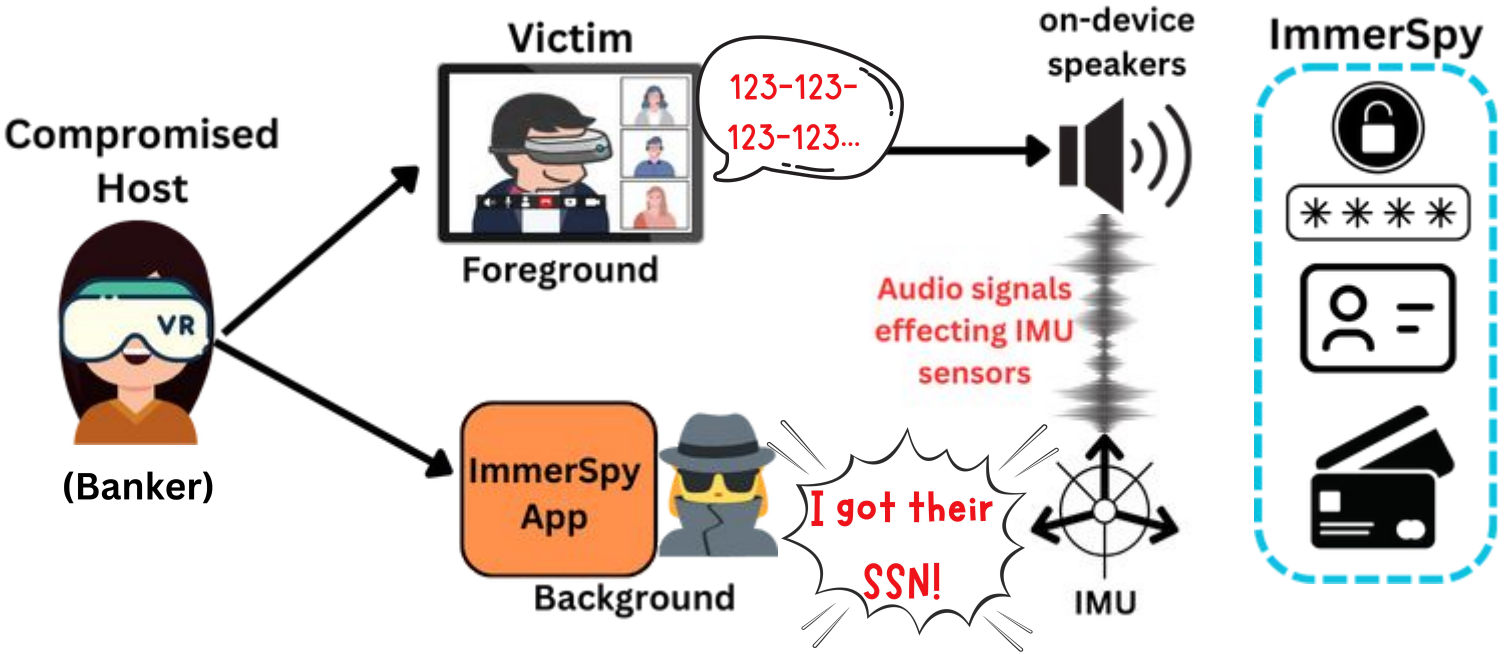


Threat model



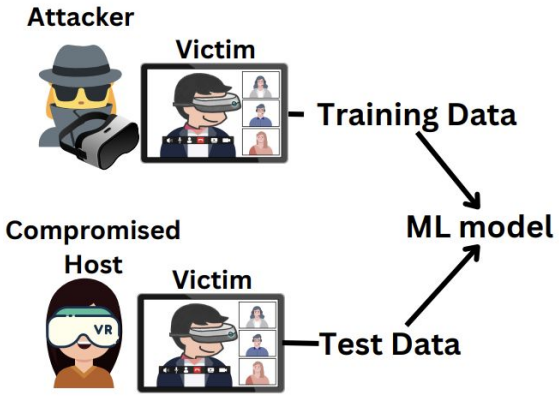
Threat model

Key insight: A malicious app can exploit the accelerometer data to recover the spoken digit content of the users, i.e., date of birth, credit card numbers, SSN.

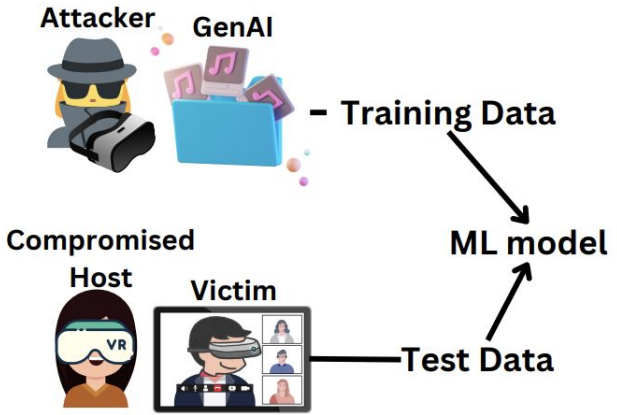


Attacker scenarios

1. The attacker may either have the labeled data of the victim **(Informed Attacker)**
2. Or they may operate fully black-box **(Uninformed Attacker)**
3. They can use **GenAI TTS** models to form their training dataset



Informed Attacker
(labeled data of the victim)



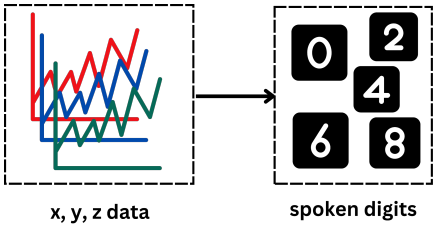
Uninformed Attacker
(black-box attack, no labeled data of the victim)

VR specific challenges for the adversary

C1) Head-associated movements on the sensors.

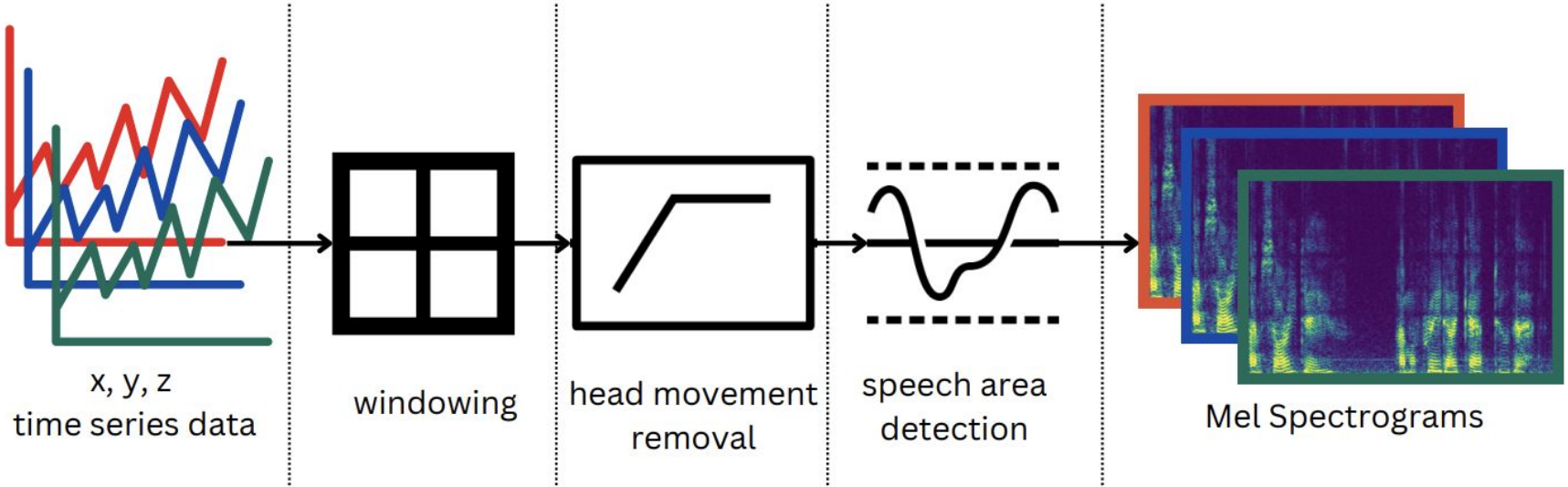
C2) Low sampling rates for speech spectrum.

C3) Extracting audio-signal characteristics from x, y, z values.



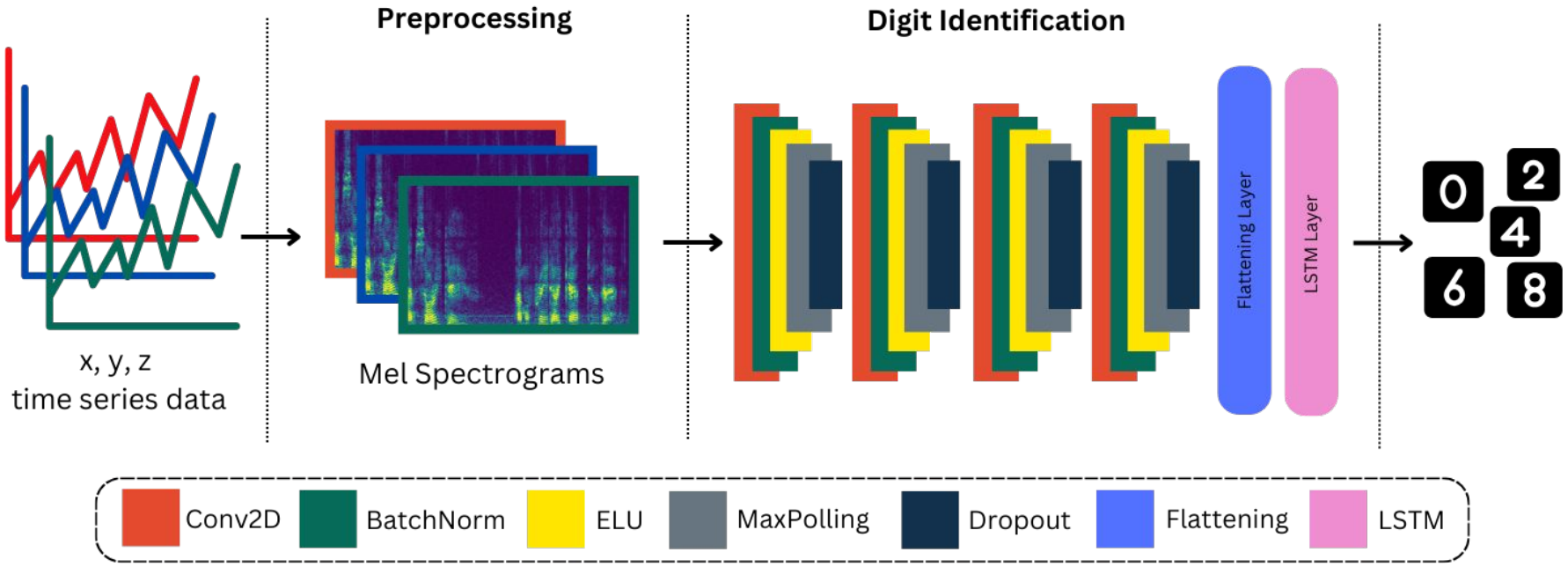
ImmerSpy overview

We present **ImmerSpy** which is a novel speech inference attack that leverages zero-permission sensors on VR devices.



ImmerSpy overview (cont'd)

We leverege a Mel spectrogram-based CNN-LSTM network that capture temporal dependencies and dynamics within speech signals.



Evaluation

Collected accelerometer data from Meta Quest 2 and 3 while speakers are playing audio:

- Presence of speech and without
- Online audio files of pronounced digits in English
- Head movements included
- Total ~6 hours of data (i.e., 1000 samples/sec, total: 21M samples)
- Used open-source GenAI TTS models with 80 AI-voices



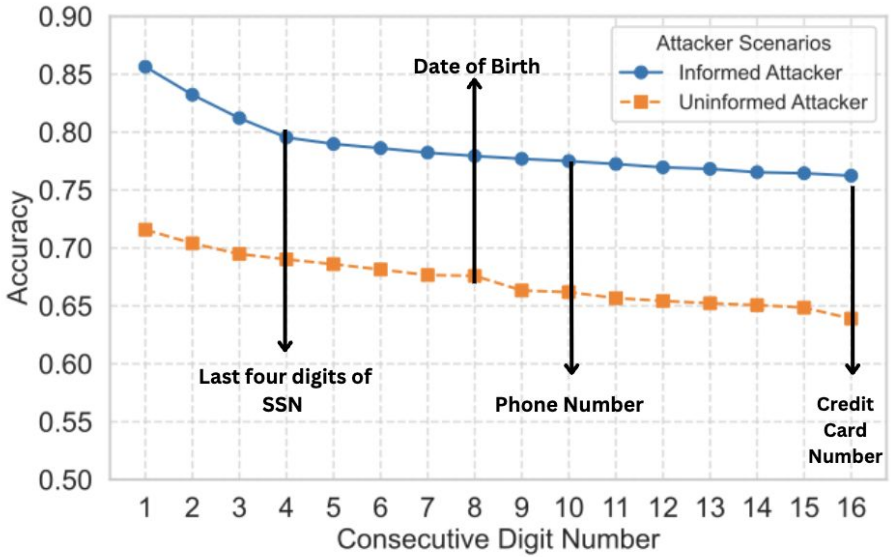
ImmerSpy's effectiveness

Compared it with baseline models and evaluated the accuracy of recovering consecutive digits.

ImmerSpy correctly guesses spoken digits by:

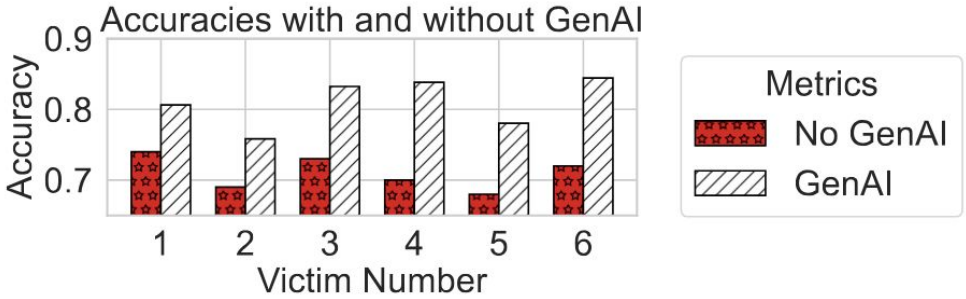
- >85% for Informed attack
- >72% for Uninformed (black-box) attack
- Top-3 guesses reaches up to 99%

ImmerSpy is highly effective.



Effect of using GenAI

We also enhanced the performance of the attack through expanding our dataset by adding 80 different AI-generated voices.

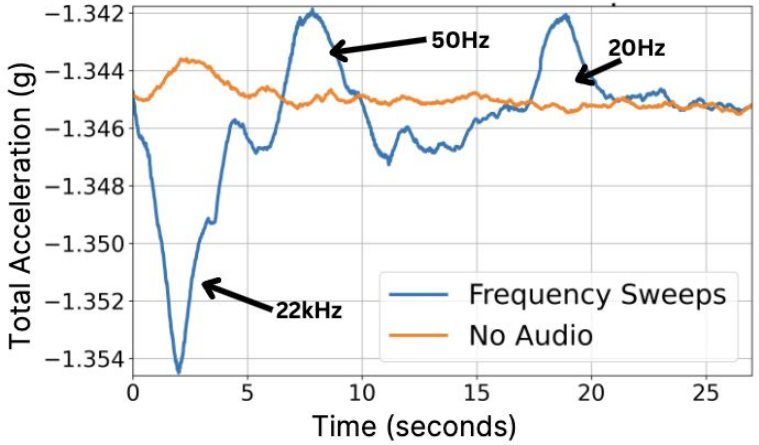


When the attacker lacks their victim's data, they can enhance their dataset with utilizing AI-generated voices.

Black-box attacker's accuracy increases to 82% from 71%.

Countermeasure

We propose a novel defense which takes advantage of the futures of ImmerSpy by introducing noise to the sensor data using inaudible sounds.



	Informed attacker	Uninformed attacker
Attack	85.6%	71.5%
With Defense	9.3%	9.6%

Our defense solution drastically reduces the attack's accuracy!

Concluding remarks

ImmerSpy is a concerning speech inference attack on VR devices via accelerometer data.

- Captures spoken digits with high accuracy even in black-box scenarios.
- Uses GenAI to improve the attack performance
- Uses inaudible sounds as a defense

References

[1]Derin Cayir, Reham Mohamed, Riccardo Lazzeretti, Marco Angelini, Abbas Acar, Mauro Conti, Z Berkay Celik, and Selcuk Uluagac. Speak up, i'm listening: Extracting speech from zero-permission vr sensors. In Proceedings of the Network and Distributed System Security Symposium (NDSS), 2025.

[2] D. Cayir, A. Acar, R. Lazzeretti, M. Angelini, M. Conti, and S. Uluagac, "Augmenting security and privacy in the virtual realm: An analysis of extended reality devices," IEEE Security & Privacy, 2023 (IEEE Security and Privacy Magazine Best Paper Award).

[3] A. K. Sikder, H. Aksu, and A. S. Uluagac, "6thsense: A context-aware sensor-based attack detector for smart devices," in USENIX Security Symposium, 2017.

[4]A. K. Sikder, G. Petracca, H. Aksu, T. Jaeger, and A. S. Uluagac, "A Survey on Sensor-based Threats and Attacks to Smart Devices and Applications," IEEE Communications Surveys and Tutorials, 2021.

[5]H. Farrukh, R. Mohamed, A. Nare, A. Bianchi, and Z. B. Celik, "{LocIn}: Inferring semantic location from spatial maps in mixed reality," in USENIX Security Symposium, 2023.

[6]Z. Ling, Z. Li, C. Chen, J. Luo, W. Yu, and X. Fu, "I know what you enter on gear vr," in IEEE Conference on Communications and Network Security (CNS), 2019.

[7]V. Nair, W. Guo, J. Mattern, R. Wang, J. F. O'Brien, L. Rosenberg, and D. Song, "Unique identification of 50,000+ virtual reality users from head & hand motion data," arXiv preprint arXiv:2302.08927, 2023.

[8]T. Zhang, Z. Ye, A. T. Mahdad, M. M. R. R. Akanda, C. Shi, Y. Wang, N. Saxena, and Y. Chen, "Facereader: Unobtrusively mining vital signs and vital sign embedded sensitive info via ar/vr motion sensors," in ACM SIGSAC Conference on Computer and Communications Security, 2023.

[9]Y. Zhang, C. Slocum, J. Chen, and N. Abu-Ghazaleh, "It's all in your head(set): Side-channel attacks on AR/VR systems," in USENIX Security Symposium, 2023.

[10]Y. Michalevsky, D. Boneh, and G. Nakibly, "Gyrophone: Recognizing speech from gyroscope signals," in USENIX Security Symposium, 2014.

Thank you! Questions?

Derin Cayir

dcayi001@fiu.edu

CSL Lab

We would like to thank our sponsors and anonymous reviewers. And, we greatly appreciate the funding for this work. The views expressed are those of the authors only, not of the funding agencies.