F²Key: Dynamically Converting Your Face into a Private Key Based on COTS Headphones for Reliable Voice Interaction

Author: **Di Duan**, Zehua Sun, Tao Ni, Shuaicheng Li, Xiaohua Jia, Weitao Xu, Tianxing Li

Presenter: Di Duan (duandiacademic@gmail.com)

City University of Hong Kong

Michigan State University

香港城市大學
City University of Hong Kong

Teaser Trailer

[Fake] An interview video with Elon Musk during a TED talk.

Threat Model & Motivation

Threat Model

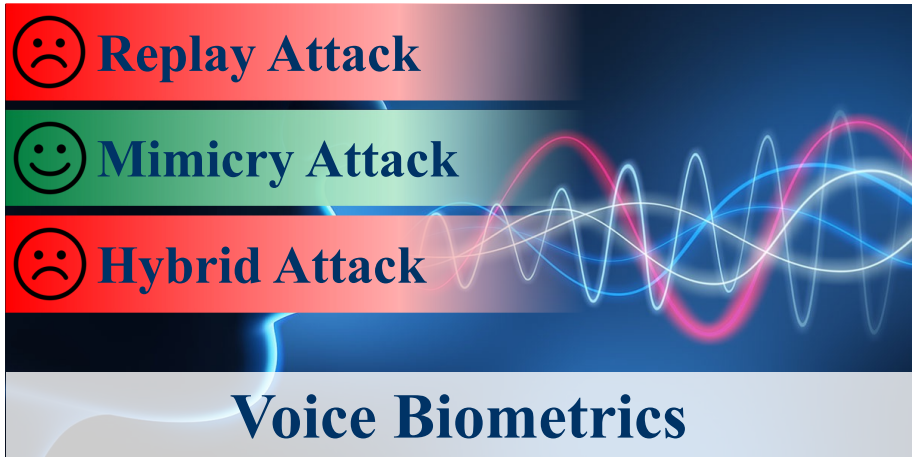
Threats in Speech-Included Artifacts

Replay Attack, Mimicry Attack, Hybrid Attack …

Synthesis Attack (Deepfake, Voice Cloning)

Existing Solutions

Combat Speech-Involved Attacks



Existing Solutions

Recent New Insight

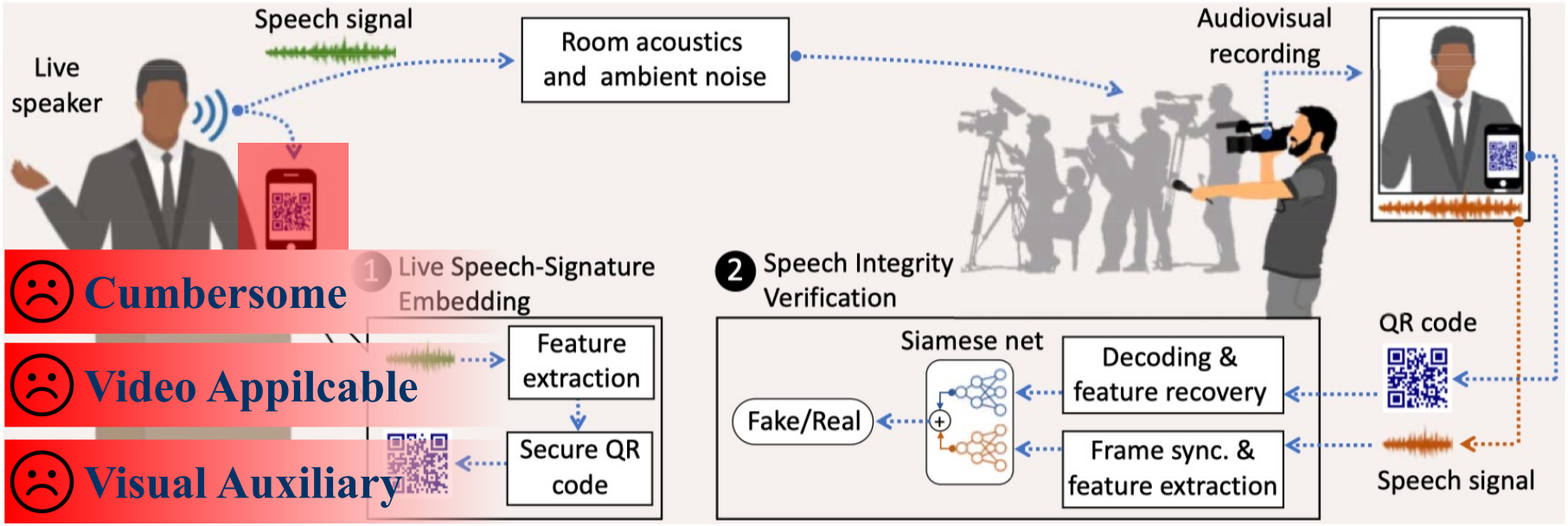


Figure 1: Design overview: ❶ The live speech-signature embedding module extracts features from speech signals in real-time and generates a sequence of cryptographically secure QR codes. ❷ The speech integrity verification module uses our algorithm to check the speech in the content under question matches with the features recovered from the QR codes visible in the video.

[MobiSys'23] *"Is this my president speaking?" Tamper-proofing Speech in Live Recordings*

Motivation

A Question

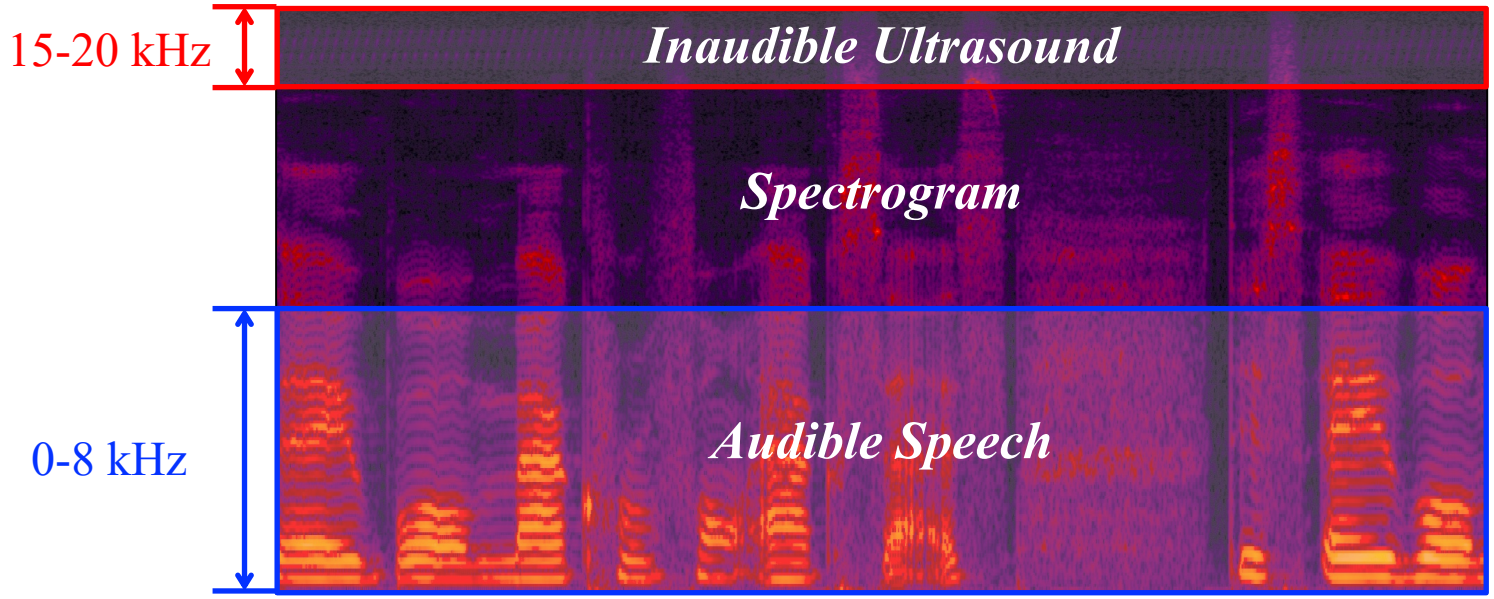
Is a reliable non-visual-aided replay-attack-resistant continuous user-friendly mimicry-attack-resistant hybrid-attack-resistant pure-audio-applicable **solution feasible?**

Our Idea

Embedding Non-Visual Physical Information from the Real-World!

Idea

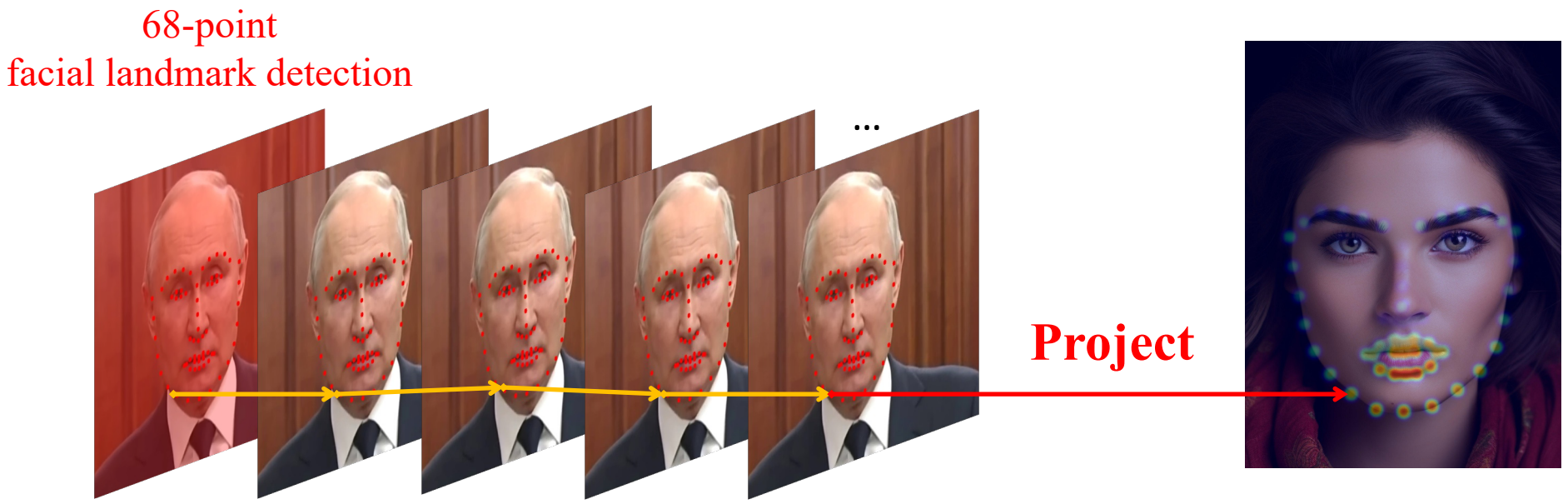
Which Modality?



Ultrasound should be an excellent choice!

Idea

Which areas are active during speech?

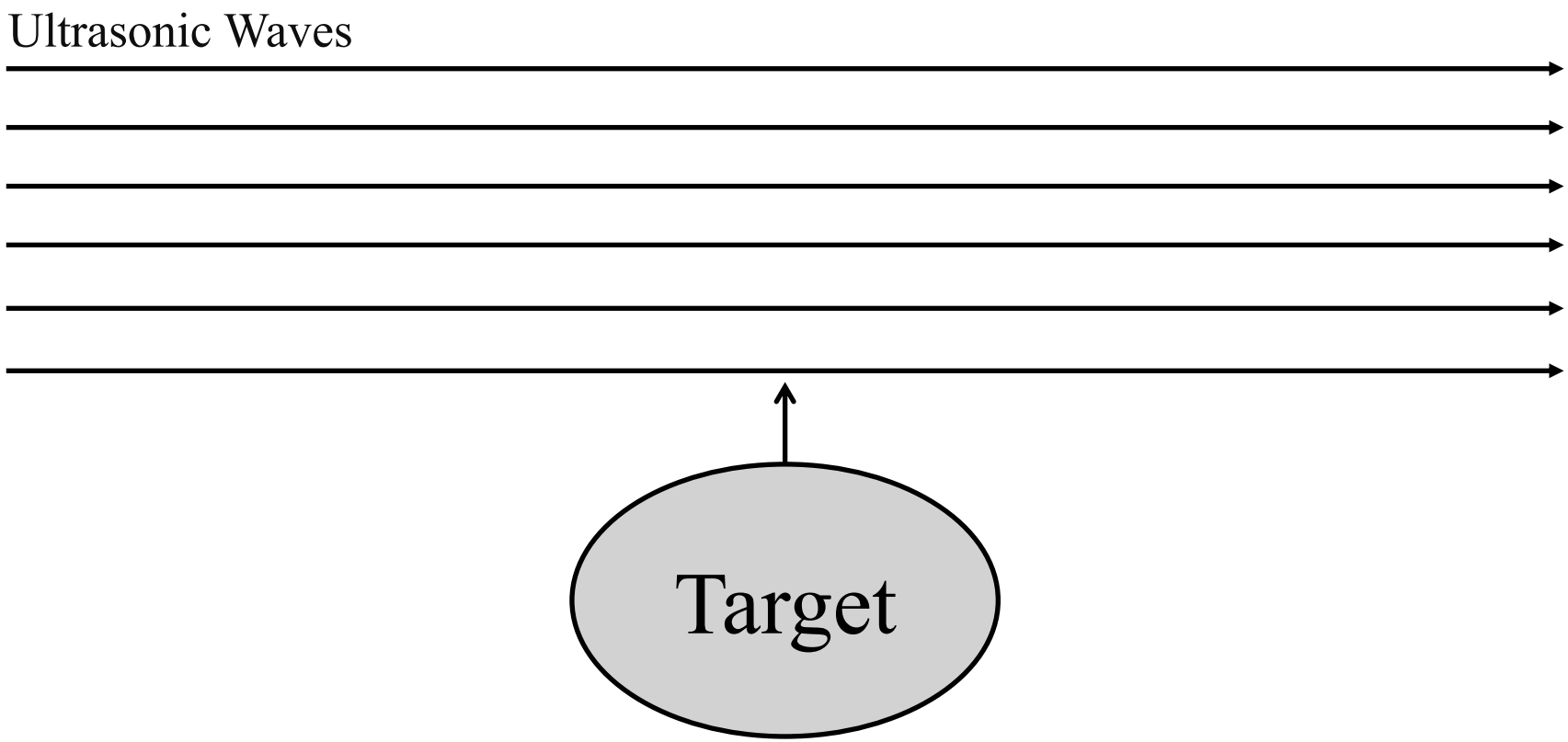


Analysis on single-person TV address video.

Active areas.

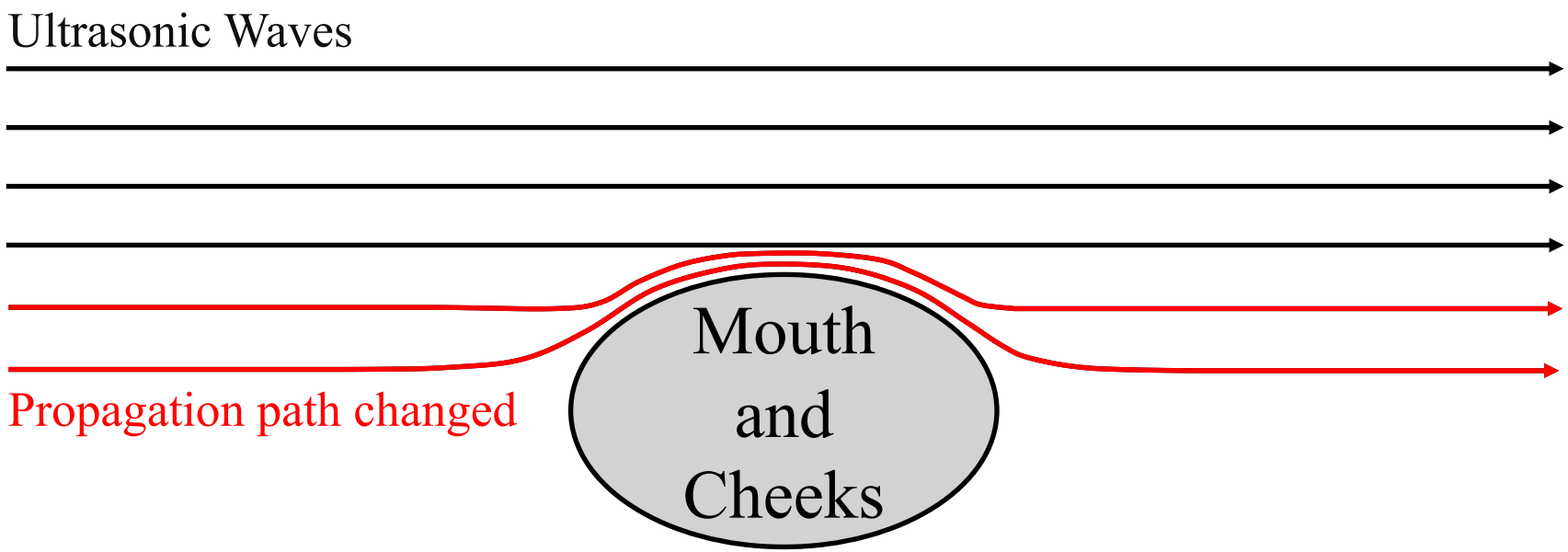
Idea

Articulatory Gestures Detection



Idea

Articulatory Gestures Detection



The propagation path changes can be reflected in Channel Impulse Response (CIR) profiles.

Idea

Sensing Platform

Hands-Free, Head-Mounted

Continuous Sensing

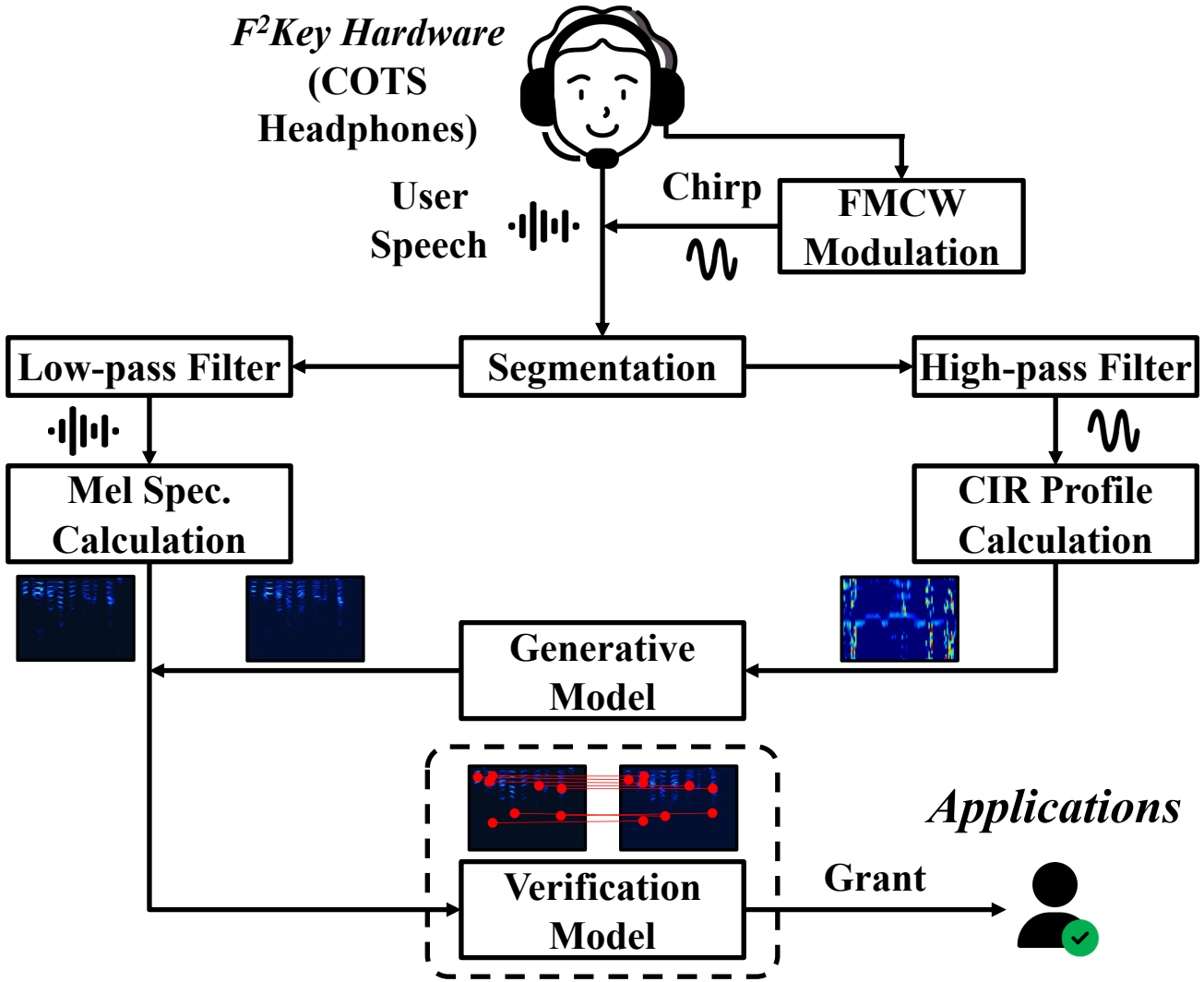
Created by Midjourney

Portable, Ubiquitous, User-friendly

Rich Information

System Overview

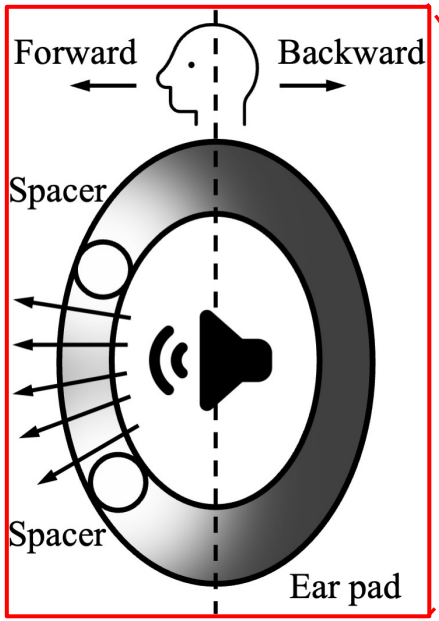
System Overview

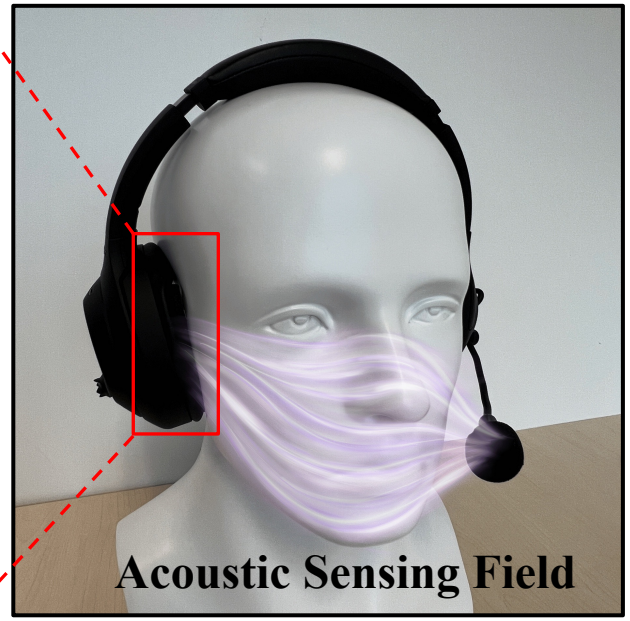


Challenge #1:
Extremely Low SNR of Escaped Ultrasound

Method #1

Create A Gap for Ultrasound to “Escape”

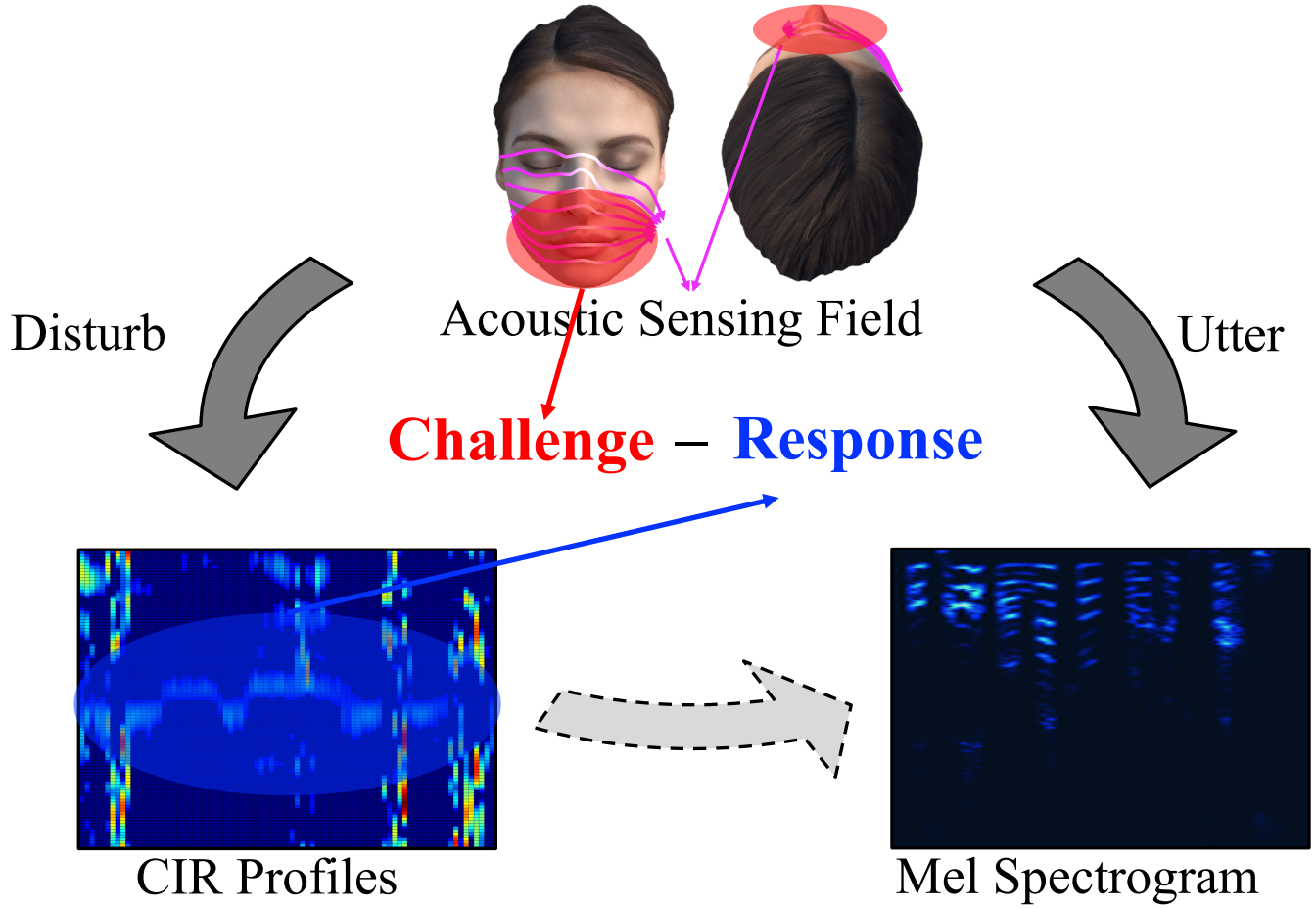




Challenge #2:
The Ambiguous Relationships Between Acoustic Features and Speech

Method #2

Challenge-Response Mechanism



Challenge #3:
Dynamic Verification Rather than Comparing with A Static, Fixed, Pre-Established “Template”

Method #3

A Key Insight

Dynamic verification of user legitimacy is possible by embedding the unique CIR-spectrogram mapping relationship of each individual into a user-specific generative model

User-Specific Model

$$S = F(C; \theta, \phi)$$

Spectrogram

CIR Profiles

Facial Structure

Private key (hard to be stolen)

Evaluation

Evaluation

Data Collection, Experimental Setup, and Evaluation Metrics

Data Collection

- 26 participants (14 males, 12 females), age 19 to 35.
- 4×4 meters room
- 15 security-related sentences, each one is repeated 30 times
- 11700 utterances in basic dataset

Experimental Setup

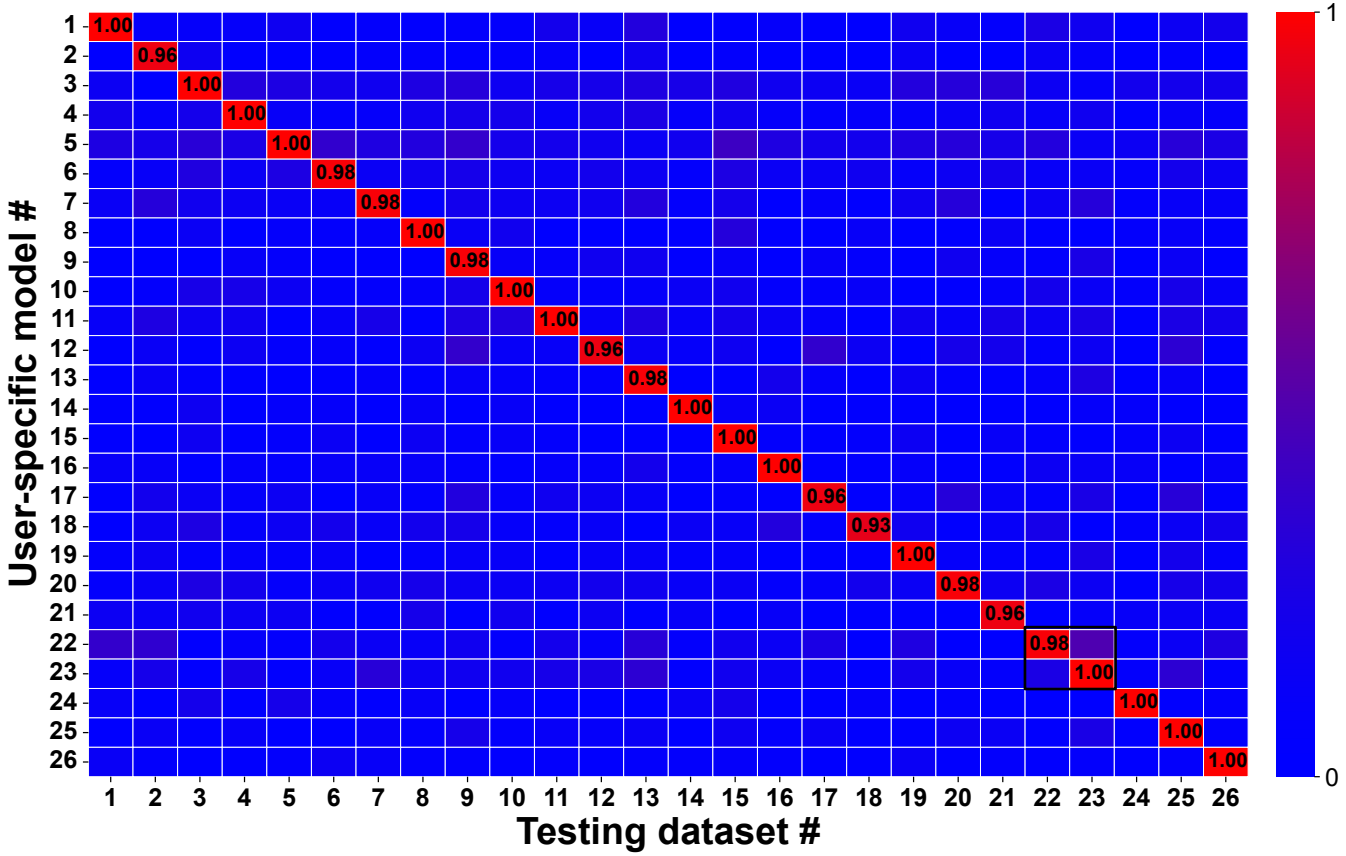
- Three headphone model: Sony WH-1000XM4+Antlion Mod Mic, Logitech G733, ATH-G1WL
- Workstation: AMD Ryzen 3955WX, 4 × 64 GB of RAM, single NVIDIA RTX 3090 GPU
- PyTorch 1.13

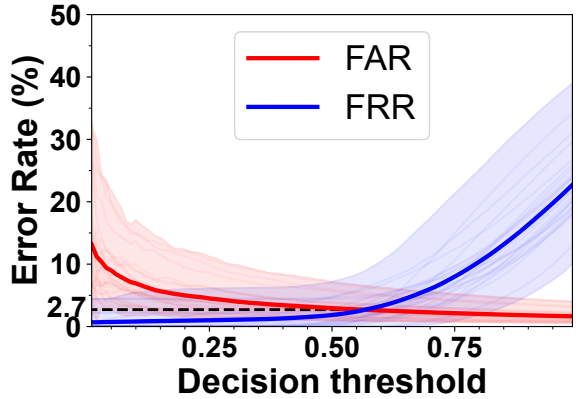
Evaluation Metrics

- True Acceptance Rate, False Acceptance Rate, False Rejection Rate, Equal Error Rate

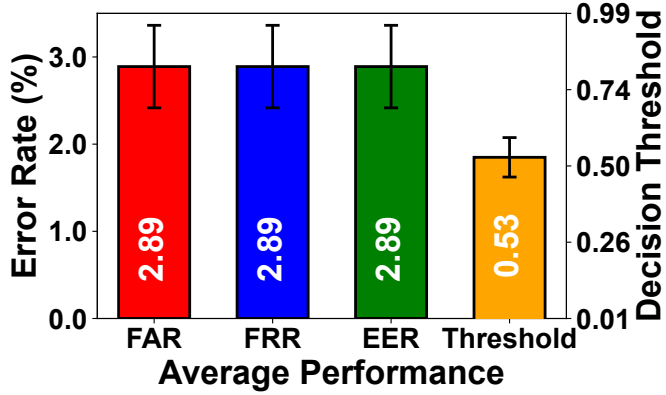
Evaluation

Zero-effort Attack





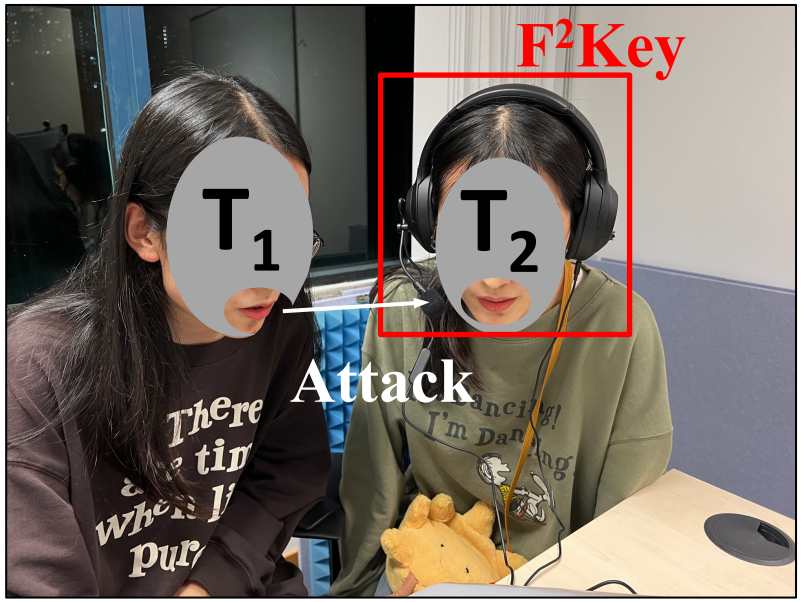
Average FAR FRR and EER among 26 participants

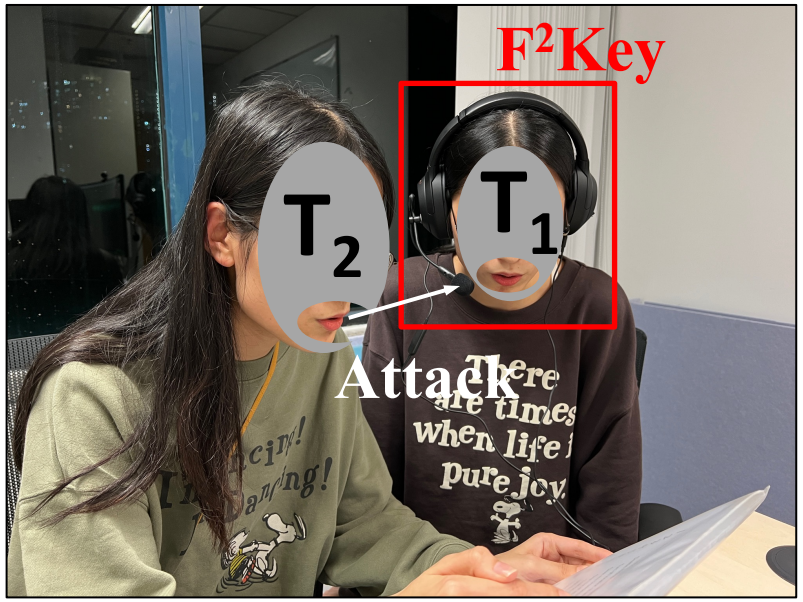


Each participant's FAR FRR and EER

Evaluation

Identical Twins Study

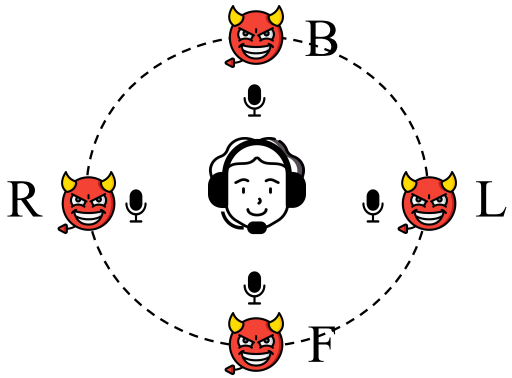




Identical twins T_1 and T_2 attack each other, while the victim performs silent speech to provide CIR profiles. The attack success rate was 24.5% on average.

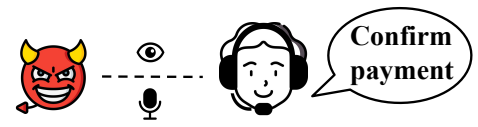
Evaluation

Replay/Mimicry/Hybrid Attack

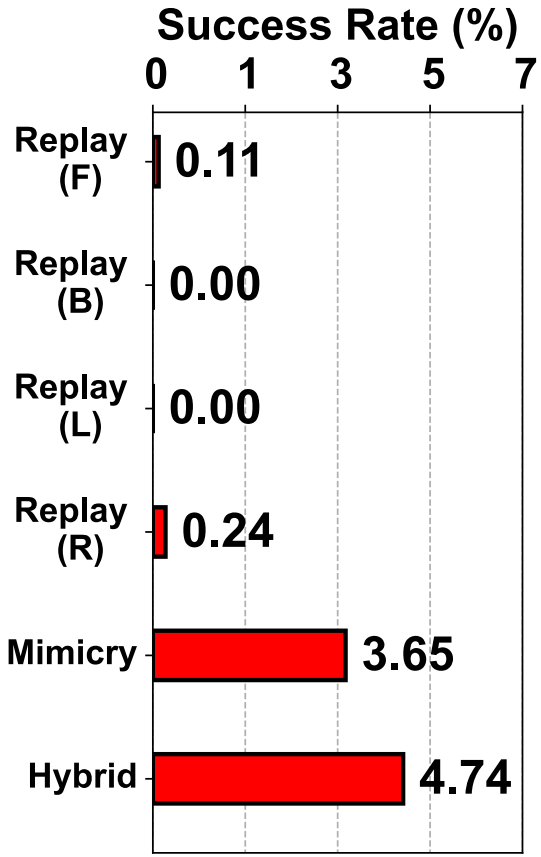


Replay Attack

Mimicry Attack



Hybrid Attack



Conclusion

- We proposed the first earable physical security system that embeds physical information into speech for anti-counterfeiting artifacts.

- We addressed three main challenges to realize our system by designing a new hardware setting, modelling the relationship between articulatory gestures and user speech, and embedding the mapping relationship into a generative model.

- Our evaluation demonstrates that F^2Key can resist over 95% of various attacks, such as replay attack, mimicry attack, and hybrid attack.

Thanks for your attention!

Email: duandiacademic@gmail.com

Personal Homepage