Turning Everyday Earphones into a Full-Duplex Speech Eavesdropping Platform via Zero-permission IMU

Ming Gao
Nanjing University of Posts and Telecommunications
Nanjing, China
gaomingppm@njupt.edu.cn

Ayijiaken Amantai
Nanjing University of Posts and Telecommunications
Nanjing, China
2024040503@njupt.edu.cn

Yichen Dai
Nanjing University of Posts and Telecommunicationsa
Nanjing, China
yichen.dai@njupt.edu.cn

Jia Lv
Nanjing University of Posts and Telecommunications
Nanjing, China
1025040719@njupt.edu.cn

Kaiyan Cui
Nanjing University of Posts and Telecommunications
Nanjing, China
kaiyan.cui@njupt.edu.cn

Jinsong Han
Zhejiang University
Hangzhou, China
hanjinsong@zju.edu.cn

Minhao Cui
Seoul National University
Seoul, Republic of Korea
minhaocui@snu.ac.kr

Fu Xiao*
Nanjing University of Posts and Telecommunications
Nanjing, China
xiaof@njupt.edu.cn

Abstract

Speech eavesdropping has become a significant security concern. While non-acoustic side channels (e.g., mmWave and RFID) offer eavesdropping pathways, they invariably rely on external hardware deployed in proximity to victims. This limits their stealth and practicality. Earphones, by contrast, pose a more insidious threat. Their embedded inertial measurement units (IMUs), accessible without user permissions, create a hardware-free stealthy side channel. In this paper, we propose a novel side-channel attack that achieves the first open-vocabulary, full-duplex speech eavesdropping on commercial off-the-shelf (COTS) earphones. This attack uniquely accomplishes full-duplex recognition of both 'what you say' (user speech) and 'what you hear'(audio played through the earphones), even when overlapping (e.g., talking while listening to music or multi-person conversing). We resolve the severe mutual interference inherent to this underdetermined signal separation problem. Constrained by COTS earphones' fixed IMU sampling rates of merely 25 Hz, we propose a multi-axial data fusion strategy to extract speech features from this limited bandwidth. A dual-stream attention network, along with a phoneme-based pixel-reorganization strategy (to support open-vocabulary recognition) and large language model (LLM)-assisted refinement, is employed to accurately recognize sensitive speech. Extensive evaluations on COTS earphones demonstrate the attack's practicality and urgent threat to speech privacy.

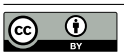
CCS Concepts

• **Security and privacy → Mobile and wireless security**.

Keywords

Side-channel attack, full-duplex eavesdropping, earphone, IMU

*Fu Xiao is the corresponding author.



CCS '26, The Hague, Netherlands

ACM ISBN 979-8-4007-2871-6/2026/11
https://doi.org/10.1145/3830454.3832674

ACM Reference Format:
Ming Gao, Ayijiaken Amantai, Yichen Dai, Jia Lv, Kaiyan Cui, Jinsong Han, Minhao Cui, and Fu Xiao. 2026. Turning Everyday Earphones into a Full-Duplex Speech Eavesdropping Platform via Zero-permission IMU. In *Proceedings of the 2026 ACM SIGSAC Conference on Computer and Communications Security (CCS '26), November 15–19, 2026, The Hague, Netherlands.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3830454.3832674

1 Introduction

Speech privacy has long been a fundamental security concern in society. Conversations frequently contain highly sensitive personal and financial information, making them prime targets for malicious actors. With the proliferation of electronic devices around us, the risk has become much higher. Fortunately, directly capturing the voice information from acoustic sensors (i.e., microphones) is now often strictly protected [1–6]. Thus, adversaries are shifting toward side-channel attacks that exploit non-acoustic sensors.

Speech, produced by a human voice or a loudspeaker, propagates as mechanical waves that induce subtle vibrations in surrounding objects. As these vibrations are inherently and passively generated during speech, they establish a reliable, always-on side channel that effectively bypasses existing microphone-centric security defenses [1–5]. Adversaries can exploit this channel via various sensing technologies, e.g., mmWave [7–12], optical [13, 14], RFID [15–17], and UWB [18, 19], to capture the vibrations and thereby eavesdrop on a victim's speech. However, these existing side-channel attacks face a practical limitation. They rely on additional, often obtrusive, hardware setups, such as RFID readers or radar devices, which must be physically positioned adjacent to the victim [7–19]. This requirement severely restricts their real-world applicability and threat potential.

Given the properties of such side channels, we wonder: *Are there readily available non-acoustic sensors near a victim that could be exploited for eavesdropping?* To answer this question, we identify wireless earphones, with over 3 billion units shipped globally each year [20], as a widespread yet overlooked attack surface. Despite their most sensitive components (microphones and speakers) being protected by stringent permission mechanisms [1], these earphones are far from secure. We explore that the built-in inertial measurement units (IMUs) offer a novel side channel, which is increasingly

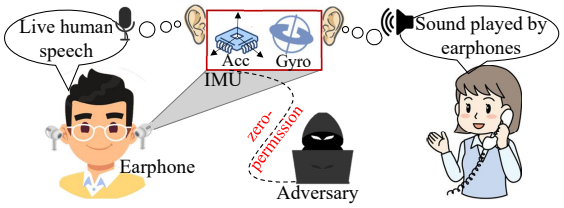


Figure 1: Remote zero-permission access to IMUs enables full-duplex eavesdropping on both microphone-received speech and speaker-produced audio of COTS earphones.

embedded in commercial off-the-shelf (COTS) earphones (e.g., Apple AirPods, Samsung Galaxy Buds) [21, 22] to support functions like head tracking [22–24]. This IMU-based channel exhibits two key characteristics that highlight its threat potential:

- **Zero permission:** Regarded as low-risk and privacy-benign, IMUs allow third-party apps or URLs to access raw data streams continuously in the background without user consent or notification [21, 25, 26]. This empowers remote adversaries to maliciously utilize the victim's own earphones as spy tools, eliminating the need for external hardware or physical proximity.
- **Duplex capability:** Earphones hold a unique position as a bidirectional audio convergence point. Therefore, their built-in IMUs can capture vibrations from the wearer's live speech, as well as from audio played by the earphone speakers. This enables unprecedented duplex eavesdropping, different from existing side channels that primarily capture simplex signals.

As illustrated in Figure 1, these characteristics expose a novel vulnerability. Adversaries can leverage the IMU side channel to monitor both 'what you say' and 'what you hear' without accessing the microphone or speaker stream.

However, in practice, implementing IMU-based side-channel attacks for duplex eavesdropping faces three unresolved challenges:

1) *Low sampling rates.* Due to COTS earphones' inherent constraints on size and battery life, their built-in IMUs operate at an unalterable low sampling rate of 25 Hz [27–29]. By the Nyquist theorem, this yields a narrow bandwidth of only 12.5 Hz, which falls well below the 85∼255 Hz fundamental frequency range of human voices [2]. Such bandwidth limitation induces heavy information loss, impairing speech semantic extraction from undersampled data.

2) *Channel interference.* Interference caused by a user's body movements can significantly distort IMU side channels [25, 30]. Even worse, the duplex capability is a double-edged sword: human-voiced speech and earphone-generated sounds may occur simultaneously, causing their induced vibrations to overlap in the IMU side channel and interfere with each other. Using earable IMUs alone, it remains an underdetermined question to separate the overlapped characteristics of dual sources in a motion-robust manner.

3) *Poor generalization.* It is impractical for adversaries to collect large-scale labeled datasets spanning diverse individuals. Consequently, existing side-channel attacks are largely confined to closed-vocabulary settings or specific user groups, which severely limits their generalization. How to achieve real-world eavesdropping that generalizes to arbitrary victims and open-vocabulary speech remains an open challenge.

We address these challenges for the zero-permission side channel by advancing undersampled signal recognition, underdetermined signal separation, and real-world attack generalization. To overcome the 25 Hz sampling-rate limitation, we exploit the rich spatial information available in IMU signals to compensate for the limited temporal resolution. Specifically, we fuse multi-axis data from accelerometers and gyroscopes using a dual-stream neural architecture with attention mechanisms, which mitigates information loss caused by the narrow eavesdropping bandwidth and enables effective speech recovery. To mitigate interference in the side channel, we observe that motion artifacts exhibit more limited patterns than speech-induced signals and accordingly suppress them via regression. We further identify a key insight: gyroscopes are significantly more sensitive to jaw-induced speech vibrations than to loudspeaker-generated vibrations. Leveraging this, we propose an iterative strategy to separate human speech from earphone-generated sounds. To extend the attack generalizability, we exploit the fact that the phoneme inventory of a language is far smaller than its vocabulary [31] through phoneme-based pixel reorganization for open-vocabulary scalability. Using this controlled dataset, an unsupervised domain adaptation framework enables cross-individual generalization. Besides, our LLM-assisted contextual refinement further amplifies the attack's real-world threat potential.

By combining the above efforts, we facilitate full-duplex rather than mere half-duplex eavesdropping, the latter being limited to capturing only one direction (either live speech or played audio) at any given moment. Our attack, however, can simultaneously decode both human-voiced and earphone-produced signals. This full-duplex capability ensures comprehensive surveillance in real-world complex scenarios. For instance, when a victim engages in dual interactions, such as listening to music or a voice assistant's response via earphones while concurrently conversing with nearby individuals [32, 33], we can capture the complete interaction context without losing information from either speech stream. In our experiment involving 52 participants across 10 popular COTS earphone models (including 7 earbuds and 3 headsets from three brands), sentence recognition accuracy reaches over 99%. To mitigate this threat, we design a defensive strategy.

Our contributions are summarized as follows.

- We propose a novel side-channel attack that remotely eavesdrops on speech by leveraging the zero-permission built-in IMUs in COTS earphones. Unlike existing approaches that rely on externally deployed devices, our method operates more covertly with fewer detectable anomalies, posing a heightened privacy threat.
- We mitigate the resolution limitation imposed by the 25 Hz low sampling rate via multi-axial IMU data fusion. Based on a rigorous theoretical model, our multi-channel strategy breaks through undersampling restrictions in side-channel signals in a software-only manner, without hardware upgrades.
- We present a novel approach for eavesdropping of bidirectional earphone audio. To our knowledge, this work achieves the first remote full-duplex eavesdropping on mobiles, which presents a far more comprehensive threat than simplex eavesdropping.
- Through comprehensive experiments, our proposed attacks are demonstrated to be effective across 10 popular COTS earphones in the real world. Experimentally, nearly all sentences spoken via earphones by 52 participants are successfully recognized.

2 Background and Threat Model

2.1 Prevalence of IMUs in Earables

Recent years have witnessed a rapid evolution of earphones into smart 'earables'. To support advanced functions like head-tracking and spatial audio [22–24], a growing number of COTS earphones, including Apple AirPods, Samsung Galaxy Buds, and Sony WF series, are equipped with built-in IMUs [21, 22]. They detect the user's head motion (via 3-axis accelerometers) and orientation (via 3-axis gyroscopes).

However, the widespread deployment of these sensors introduces a new attack surface. Unlike microphones, which are protected by strict permission systems [1], IMU data can often be accessed by third-party applications with zero permissions [21, 25, 26], making them an ideal covert channel for eavesdropping.

It has been reported that IMUs (with sampling rates over 200 Hz) can inadvertently capture subtle vibrations correlated with speech [25, 26, 34, 35] and audio [30, 36–43]. However, these findings have primarily focused on scenarios involving smartphones.The unique and pervasive context of consumer earables remains largely unexplored for eavesdropping purposes. As we detail in Section 3.2, the same physical principles apply, yet the distinct challenges and opportunities of this platform present a novel security gap.

2.2 Threat Model

2.2.1 Attack Scenario. We consider a victim wearing IMU-equipped COTS earphones connected to a smartphone. The victim engages in activities that involve both speaking (e.g., voice commands, conversation) and listening (e.g., music, calls, notification sounds).

The attack explicitly targets earphones equipped with built-in IMUs. While not present in ultra-low-cost models, these sensors are standard in mid-to-high-end devices, whose market share is increasingly growing, e.g., Apple AirPods, Samsung Galaxy Buds [21, 22].

2.2.2 Attacker Capabilities. The attacker is assumed to have tricked the victim into installing a seemingly benign application (e.g., a fitness tracker or a game) on the smartphone, following the identical attacker capabilities in [25, 26, 30, 35–43].

- **Zero-Permission Access:** Once installed, the application covertly runs in the background, collecting accelerometer and gyroscope data from earphones via standard zero-permission APIs [21]. The collected data is then used for eavesdropping model training. Additionally, the attacker can identify the earphone model by reading the publicly accessible device specifications, which are typically neither sensitive nor protected.
- **Remote Execution:** Unlike existing side-channel attacks that require physical proximity and specialized hardware (e.g., RFID readers placed near the victim [44]), our approach is entirely software-based and can be executed remotely.
- **Sampling Rate Limitation:** Our spyware application captures IMU readings at 25 Hz. This default inertial sampling rate of 25 Hz in COTS earphones is not allowed to be modified [27–29] due to constraints in device miniaturization and battery longevity.

2.2.3 Attack Surface Expanding: From Simplex to Full-Duplex. Earphones occupy an irreplaceable position as an intrinsic bidirectional audio duplex node. Their built-in IMUs inherently capture two distinct mechanical vibration sources that form a duplex side channel:

- **Earphone-produced audio:** Vibration from the speaker diaphragm during audio playback (e.g., music, calls, notifications).
- **Human-voiced speech:** Vibration from the wearer's vocal cords and facial movements during speaking.

This unique dual-channel attribute enables potential full-duplex acoustic eavesdropping, distinguishing it from conventional simplex side-channel attacks [25, 26, 30, 35–43].

Real-world earphone usage is predominantly full-duplex, as users frequently engage in concurrent audio input and output behaviors. For instance, a user may speak to nearby persons or issue voice commands to a smart assistant while listening to music, watching videos, or receiving continuous audio feedback (e.g., turn-by-turn navigation, game sound effects) [32, 33]. Similar full-duplex scenarios are also common in teleconferencing, where a participant might speak while other attendees are talking or while background notification alerts are playing. While we acknowledge that strict simultaneous conversations are relatively rare in formal calls, mutual interference between the user's speech and the earphone's output remains ubiquitous in broader usage scenarios. These emphasize the practical relevance of full-duplex eavesdropping on earphones.

Nevertheless, existing acoustic eavesdropping attacks typically assume a simplex [25, 26, 30, 35–43] scenario: either capturing speech in quiet or only speaker output. A recent work [44] leverages RFID tags for half-duplex eavesdropping. However, it requires malicious peripherals with physical contact, i.e., attaching the tags to victim devices, with RFID readers and antennas positioned nearby. These requirements limit its practical threat. In short, prior work has overlooked full-duplex acoustic characteristics, leaving a significant blind spot in acoustic side-channel risk.

To fill this gap, we build a versatile attack that is robust enough to handle complex full-duplex eavesdropping while maintaining high performance in traditional simplex scenarios. By solving the hardest case (simultaneous signals), we naturally cover the easier cases (turn-taking), thereby significantly expanding the attack's real-world applicability.

3 Model and Analysis

We model the duplex side channel that maps acoustic signals to low-rate IMU measurements (25 Hz), followed by a feasibility study.

3.1 IMU Side-channel Model At 25 Hz

3.1.1 Speaker-to-IMU Side Channel. The vibrations from a speaker propagate through the solid structure of earphones, such as the shell and circuit board, and are then detected by the built-in IMUs.

Undersampling at Fs=25 Hz is considerably below the Nyquist rate required for capturing speech signals, ranging within [85, 255] Hz [2]. While pre-sampling filters are ideally supposed to eliminate these out-of-band signals, in practice, they only manage to attenuate them [38, 41, 45]. Consequently, during the analog-to-digital conversion (ADC), an acoustic signal with frequency f undergoes aliasing and distorts into a low-frequency component,

$$f^{alias} = \min |f - n \times Fs|, n \in \mathbb{N} \tag{1}$$

Despite this distortion, IMU signals still contain extractable information. The Bandpass Sampling Theorem offers a more flexible criterion [46]. Specifically, for a bandpass signal with bandwidth

B (=170 Hz here), a more relaxed condition $Fs \geq 2B$ is sufficient. Although 25 Hz still falls short, this perspective shifts the focus from the maximum frequency f_{max} to the bandwidth B.

Speech signals exhibit spectral sparsity, with phonemic information concentrated in distinct frequency components. Compressive sensing [47] leverages this sparsity to reduce the required sampling bound for recognizing such bandpass signals. The minimum sampling frequency can be approximated as

$$Fs_{min} = O(B\log(f_{max}/B)) \approx CB\log(f_{max}/B) = 119 \text{ Hz}, \quad (2)$$

where C is a constant, set as 4 typically [47]. Even though a single-axis 25 Hz stream is still below this theoretical lower bound, IMUs provide multi-axis measurements, offering six concurrent projections of the same underlying vibration source. This effectively increases the observational dimensions with more information entropy, under our basic idea of 'Space for Time'. Implementation details see Section 6, with the corresponding proof in Appendix A. Thus, aliasing merely rearranges frequencies without any information loss. It functions as a one-to-one frequency transformation.

Furthermore, while undersampling compresses the frequency, it preserves the signal's relative structure. Consider a speech signal:

$$s(t) = \sum\nolimits_k a_k e^{j2\pi f_k t + \phi_k}, \quad (3)$$

where a_k, f_k, and ϕ_k are the amplitude, frequency, and phase of the k-th component, respectively. After undersampling, the discrete-time signal becomes

$$s(nT_s) = \sum\nolimits_{k=1}^{K} a_k e^{j2\pi f_k^{alias} nT_s + \phi_k + j2\pi nT_s(f_k - f_k^{alias})}. \quad (4)$$

Since $\exists n_k \in \mathbb{N}$ such that $f_k - f_k^{alias} = n_k \cdot Fs$, the phase term is

$$2\pi nT_s(f_k - f_{Lk}) = 2\pi n/Fs \cdot n_k \cdot Fs = 2\pi n \cdot n_k,\ n \cdot n_k \in \mathbb{N} \quad (5)$$

Its phase contribution is 0 (=$2\pi n \cdot n_k \mod 2\pi$), and thus,

$$s(nT_s) = \sum\nolimits_{k=1}^{K} a_k e^{j2\pi f_{Lk} nT_s + \phi_k}. \quad (6)$$

By comparing Equation 3 to Equation 6, it is clear that the undersampled signal retains the original amplitude sequence a_k and phase relationships ϕ_k, with only the frequencies mapped downwards. Consequently, the main structural features are preserved.

3.1.2 Human-to-IMU Side Channel. Unlike the speaker-to-IMU channel, where acoustic propagation is relatively straightforward, live human speech induces side channels to earable IMUs primarily through two mechanisms: facial motion associated with articulation and bone-conducted vibrations. Both originate from vocal organs' vibrations, which can be modeled as a multi-degree-of-freedom damped system [48]:

$$m\ddot{q}_i(t) + c\dot{q}_i(t) + kq_i(t) = e^{j2\pi f_F t}, \quad (7)$$

where m, c, k represent the equivalent mass, damping, and stiffness matrices, respectively, $e^{j2\pi f_F t}$ signifies the vocal stimulus with a frequency f_F, and q_i is the i-dimensional displacement following

$$q_i(t) = Ae^{j(2\pi f_F t + \phi)},\ \dot{q}_i(t) = j\phi q_i(t), \quad (8)$$

where the gain A and the phase lag ϕ depend on f_F.

Facial movements: Speech-driven jaw and facial tissue motions, associated with displacement $q_i(t)$ [48], directly transmit to the earphone body. This causes a physical shift and rotation of the IMU. Their power spectral density $S_{mm}(f)$ is concentrated in a low-frequency band [49, 50]. We define the essential bandwidth B_ϵ as the smallest frequency satisfying

$$\frac{\int_0^{B_\epsilon} S_{mm}(f)\,df}{\int_0^{\infty} S_{mm}(f)\,df} \geq 1 - \epsilon. \quad (9)$$

For $\epsilon = 0.1$, the corresponding $B_{0.1} \leq 12$ Hz [49, 50]. Therefore, over 90% of the signal's energy resides below the Nyquist frequency. Since $q(t)$ is essentially bandlimited to $B_{0.1}$, the classic sampling theorem guarantees that $q(t)$ can be perfectly reconstructed from its samples $q[n] = q(nT_s)$, provided that $Fs > 2B_{0.1}$.

Transient events (e.g., plosive consonants) may exhibit spectral content up to $f_h \approx 20$ Hz. The energy of these components is weak. Define the folding noise ratio Γ due to aliasing as:

$$\Gamma = \frac{\int_{Fs/2}^{\infty} S_{mm}(f)\,df}{\int_0^{\infty} S_{mm}(f)\,df}. \quad (10)$$

Given the rapid spectral roll-off of $S_{mm}(f)$ above 12 Hz, Γ is negligible (< 2%), ensuring that aliasing does not corrupt the perceptually and biomechanically relevant features of $q(t)$.

Bone conduction: Vocal vibrations propagate through the cranial bones to the IMU as $\dot{q}_i(t)$. This signal follows aliasing distortion, as described by the undersampling process in Equation 1, while the multi-axis fusion would compensate for undersampling.

3.2 Feasibility Investigation

A pilot experiment is conducted on Apple AirPods Pro 2, the top-selling wireless earphones on Amazon in 2024. We measure the intensity of the built-in accelerometer and gyroscope in Table 1 under various acoustic scenarios.

Table 1: IMU response intensity in AirPods Pro 2

	Acc (m^2/s^5)	Gyro (rad^2/s^3)
Idle (no sound)	9.82×10^{-4}	1.12×10^{-5}
Earphone-produced	3.34×10^{-3}	6.21×10^{-5}
Human-voiced	1.79×10^{-3}	1.41×10^{-2}
Simultaneous	5.01×10^{-3}	1.42×10^{-2}

We first measure the intrinsic noise of the Apple AirPods Pro 2. By calling the application programming interface (API) `CMHeadphoneMotionManager.DeviceMotion Handler`, the earable IMU readings are sampled at 25 Hz by default and transmitted to an iPad 9. During this measurement, the earphones remain stationary on a desktop without playing any audio. The average power of intrinsic noise measures $9.82 \times 10^{-4}\ m^2/s^5$ in the accelerometer and $1.12 \times 10^{-5}\ rad^2/s^3$ in the gyroscope. These noise components are later removed using filters described in Section 5.1.

3.2.1 Capturing Earphone-produced Sounds. Experiments show that IMUs are sensitive to sounds generated by the earphone speakers. We play audio selected from the AudioMNIST dataset [51] through the earphones at half of the maximum volume. The IMU readings exhibit substantial changes as listed in Table 1. This is observed even though the earphones remain stationary. These fluctuations appear across all three axes of both sensors.

We play single-tone sounds at maximum volume to measure the frequency response range of the IMUs. The acoustic frequency

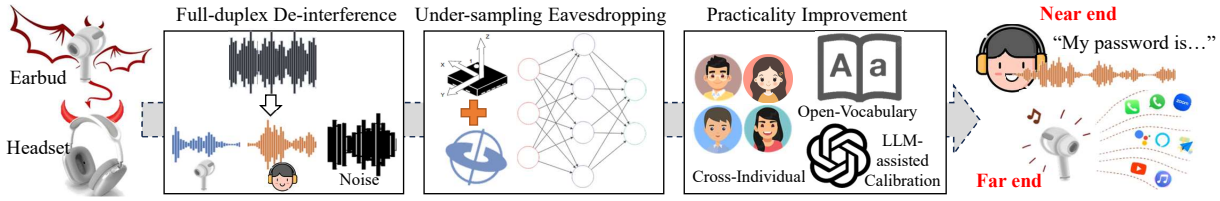


Figure 2: We propose a novel side-channel attack on COTS earphones. It can achieve full-duplex eavesdropping on both 'what you say' and 'what you hear' simultaneously. The attack is covert with less anomalous manifestations, just by leveraging built-in IMUs sampled at an extremely low rate of 25 Hz.

sweeps from 85 Hz to 255 Hz at an interval of 1 Hz. These out-of-band signals are observed to distort into a narrow band of 12.5 Hz. Both accelerometers and gyroscopes successfully capture aliased tones under the human voice's fundamental band.

3.2.2 Capturing Human-voiced Sounds. We measure the IMU response to the wearer's live speech. Two volunteers, one male and one female, wear the earphones. They remain physically still throughout the experiment. They are asked to read randomly selected English sentences at their normal speaking volumes. During this period, the earphones play no audio. Under human-voiced sounds, the accelerometer power measures $1.79 \times 10^{-3}\ m^2/s^5$ and the gyroscope power measures $1.41 \times 10^{-2}\ rad^2/s^3$. Both values are significantly higher than those when idle. This result indicates that the earable IMUs can detect human voices.

3.2.3 Simultaneous Capture. The above two phenomena can occur at the same time. The two volunteers are asked to repeat the sentences while wearing the earphones playing audio, with both actions happening simultaneously. The average intensity reaches up to $5.01\times10^{-3}\ m^2/s^5$ in the accelerometer and $1.42\times10^{-2}\ rad^2/s^3$ in the gyroscope, roughly the sum of their intensities measured individually. This implies that IMU readings integrate dual-source features simultaneously, laying a foundation for malicious speech recovery via full-duplex eavesdropping.

Multiple sensor exploitation. Existing works mainly utilize only one sensor, either an accelerometer [26, 30, 34, 39–42, 52] or a gyroscope [37]. Nevertheless, both sensors contain rich speech-related information. In particular, our pilot experiments show that the gyroscope in earable devices achieves a higher signal-to-noise ratio (SNR) than the accelerometer, which has been considered more sensitive to acoustic vibrations in traditional belief [30]. These results highlight the significance of exploiting multiple sensors comprehensively for earable sensing.

Multiple model compatibility. Similar responses are observed not only in Apple AirPods Pro 2 but also AirPods Max headsets. The built-in accelerometer achieves 12.1 dB SNR for earphone-played audio and 4.86 dB for live speech; the gyroscope reaches 1.46 dB and 29.8 dB respectively. These results confirm our attack's applicability to both COTS earbuds and headsets that hold a substantial share of the consumer earphone market.

4 Attack Overview and Formulation

We briefly introduce our proposed attack for full-duplex eavesdropping on earphones, with a problem formulation.

There remain three fundamental challenges for full-duplex eavesdropping: how to perform robust separation of mixed full-duplex signals under motion interference, effective speech semantic recovery from 25 Hz undersampled data, and scalable performance across diverse individuals and unrestricted vocabulary.

Following the mechanical wave propagation theory [53], we formally model the IMU-based side channel as

$$IMU^T = IMU^H + IMU^S + N, \tag{11}$$

where IMU^H, IMU^S, and N denote components from human-voiced speech, earphone-produced sounds, and channel noise composed mainly of motion artifacts caused by the earphone wearer's body and head, respectively. Here, the number of unknown sound sources exceeds that of observations. This thus presents an underdetermined problem of isolating both source signals IMU^H and IMU^S from a single observation IMU^T in the presence of noise N. Our objective is to recover both speech streams and extract the semantic content from each stream.

To address these challenges, we design a three-stage attack pipeline, as illustrated in Figure 2.

In full-duplex de-interference, a regression model is designed to suppress motion artifacts. Then, an iterative, generation-based module exploits the differential sensitivity of gyroscopes to isolate overlapping human-voiced speech and earphone-produced audio.

In undersampling eavesdropping, we employ the fundamental principle of 'Space for Time'. Multi-axis data from both accelerometers and gyroscopes are fused into a unified representation, followed by a dual-stream attention network to compensate for information loss caused by the low sampling rate.

To improve the applicability of our proposed attacks, we employ phoneme-based spectrogram synthesis and unsupervised domain adaptation for open-vocabulary, cross-user recognition, followed by an LLM-assisted contextual refinement to correct semantic errors.

This integrated framework enables the first zero-permission, full-duplex speech eavesdropping attack on commercial earphones.

5 Full-duplex De-interference

We perform a noise analysis and characterize the overlapping signals to facilitate single-observation separation in Equation 11.

5.1 Noise Removal

We analyze and effectively mitigate the channel noise N, which includes intrinsic noise and disturbances caused by the movement of the victim's body and head.

Intrinsic Noise. We observe that intrinsic noise is composed of a bias and Gaussian white noise. The former is removed using a mean filter, while the latter is suppressed with a Wiener filter [38].

Motion Artifacts. The victim's body and head movements result in artifacts in IMU-based sensing. These artifacts range primarily

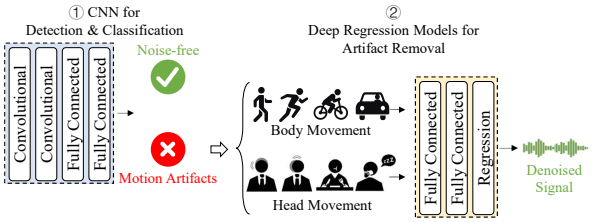


Figure 3: Our approach for artifact elimination.

below 20 Hz [26, 30]. However, according to the Nyquist sampling theorem, the inertial sensing bandwidth is limited to just 12.5 Hz. Within this narrow bandwidth, motion artifacts overlap or couple with speech-related signals. Consequently, traditional denoising approaches that rely on high-pass filtering are ineffective.

Since motion-induced interference exhibits more limited and structured patterns [25] than speech-induced signals, we propose a regression-based approach to mitigate the impact of victims' movements. The basic idea is illustrated in Figure 3. A convolutional neural network (CNN) detects whether inertial data are affected by motion artifacts and identifies the specific type of movement (i.e., body, head, or a mixture of both). Then, two deep regression models [25] are employed to compensate for artifacts caused by the body and the head, respectively. The mixed artifacts can be removed by passing through the two models successively. Both regression models share an identical architecture. They consist of two fully connected layers and a regression layer. To train these models, we collect data of several common types of movements, including 4 body movements (i.e., walking, running, bicycling, and driving) and 4 head movements (i.e., head-nodding, head-shaking, chewing, yawning). These movement data are manually mixed with speech-related readings to facilitate training.

5.2 Underdetermined Signal Separation

For effective full-duplex eavesdropping, it is essential to separate the mixed signals from both ends of the earphones.

We propose a generation-based scheme to solve the underdetermined Equation 11. We first identify a distinct difference in IMU signal intensity between human-voiced speech and earphone-generated sounds, a key observation that lays the foundation for our separation strategy. Then, we develop a mutual reconstruction mechanism that infers accelerometer signals from gyroscope measurements and vice versa. Finally, we propose an iterative strategy that integrates the intensity-based insight and mutual reconstruction mechanism, aiming to minimize cumulative errors and achieve high-precision signal separation.

5.2.1 Our Insight: Intensity Difference. Accordingly to Table 1, we observe that even though the accelerometer responses show no distinguishable differences between bidirectional sources, the gyroscope intensity during human vocalization is at least one order of magnitude higher than that during earphone-produced sounds. This distinction is intuitive, as human speech production naturally involves jaw rotation movements, to which earable gyroscopes are sensitive. In other words, the gyroscope intensity caused by human-voiced sounds accounts for the majority of the mixed signals.

Based on this observation, we initially assume that the gyroscope readings are entirely induced by human articulation. Errors

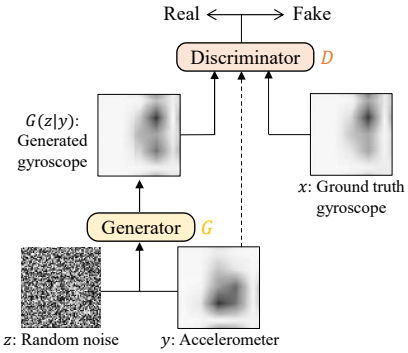


Figure 4: cGAN-based mutual reconstruction network.

introduced by this assumption are subsequently corrected through an iterative refinement process described in Section 5.2.3. Under this assumption, our next goal is to extract the earphone-generated sound component from the accelerometer data. To achieve this, we exploit the inherent correlation between accelerometer and gyroscope measurements and perform mutual reconstruction between the two sensor streams, as detailed in Section 5.2.2. Then, we can reconstruct the accelerometer component associated with human speech from the gyroscope readings and subtract it from the original accelerometer signal to isolate the earphone-induced component. Based on this isolated component, we further reconstruct the corresponding gyroscope signal from this isolated component.

5.2.2 cGAN-based Mutual Reconstruction. The accelerometer and gyroscope data represent different modalities of the same speech sources. A deterministic mapping exists between their signals, forming the theoretical basis for cross-modal synthesis. As they are recoverable from each other, generative modeling becomes a feasible approach for their mutual conversion.

We adopt a conditional generative adversarial network (cGAN) [54] to realize bidirectional transformation between the two modalities. This cGAN serves as the core tool to support our subsequent iterative separation strategy. Compared to traditional GAN [55] that generates data similar to training samples, cGAN can transform the spectrograms between different modalities [42, 56, 57] (i.e., accelerometer and gyroscope data here).

Figure 4 illustrates the architecture of our adopted cGAN. A double-threshold scheme [58] is used for speech detection and segmentation. Here, IMU readings are segmented into fixed 1-second windows, with shorter slices zero-padded. Subsequently, these segments are transformed into spectrograms via the Short-Time Fourier Transform (STFT). Taking accelerometer-to-gyroscope generation as an example, the accelerometer spectrogram acts as the condition y to generate the corresponding $G(z|y)$ based on a random noise z in the generator G. The discriminator D attempts to distinguish between the generated spectrogram $G(z|y)$ from the ground truth gyroscope spectrogram x.

In the zero-sum game of G vs D, the objective [54] is

$$\min_G \max_D V(D,G) = \mathbb{E}_x[\log D(x|y)] + \mathbb{E}_z[\log(1 - D(z|y))]. \quad (12)$$

During training, we incorporate a gradient penalty term [59] to enhance the stability of the discriminator D. This prevents the discriminator from becoming overly aggressive or encountering gradient explosions, thereby ensuring more stable training and

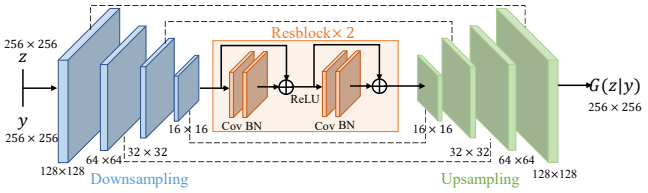


Figure 5: The architecture of our generator G: a combination of a U-Net backbone with a residual connection.

faster convergence [59]. Besides, to ensure the generated spectrograms closely match the ground truth, we implement an L_1 loss function [60] as part of our optimization objective. Hence, our final objective can be expressed as follows,

$$L = V(D,G) + \lambda_1 \mathbb{E}[(||\nabla_{G(z|y)} D(G(z|y)|y)||_2 - 1)^2] + \lambda_2 L_1, \quad (13)$$

which λ_1 and λ_2 ($\lambda_1, \lambda_2 > 0$) are the weights of the gradient penalty term and the L_1 loss term, respectively.

To address gradient vanishing while preserving channel dimensions, we modify the standard U-Net. As shown in Figure 5, a residual connection comprising two ResBlocks [61] is added between the down- and upsampling stages.

5.2.3 Iterative Source Separation. Finally, we address the errors introduced by the initial assumption in Section 5.2.1, where the gyroscope readings were assumed to originate solely from human articulation. In practice, this simplification introduces non-negligible errors, as earphone-generated sounds can also induce gyroscope responses. To mitigate these errors, we design a multi-round mutual reconstruction process, which forms the basis of our iterative separation method.

The iterative mutual reconstruction aims at accurate source separation. Initially, the mixed gyroscope data G^T is approximated as the component under human-voiced sounds G^H. With this approximation, we reconstruct the corresponding accelerometer responses A^H. Next, we subtract A^H from the mixed accelerometer data A^T via the spectral subtraction [62] to obtain the component under earphone-produced sounds A^S. From A^S, we reconstruct the corresponding gyroscope responses G^S. Finally, G^H is updated by subtracting G^S from G^T. Through repeating these steps, we successfully separate the full-duplex signals with minimal error.

Inevitably, reconstructed signals from the cGAN-based method may suffer from distortions. To alleviate the adverse impacts of such distortions, we apply a temporal averaging strategy. Specifically, for each component $IMU_n^i (IMU \in \{A, G\},\ i \in \{S, H\})$ in the n-th iteration, we replace it with the average $\frac{1}{n}\sum_n IMU_n^i$ of its counterpart from the n iterations. The iteration stops when the variation of each separated signal between two consecutive iterations falls below 1%. The detailed flow is clarified in Algorithm 1.

To support the bidirectional generation between accelerometer and gyroscope data for both human-voiced and earphone-produced sounds, we train four separate models in total. Experimentally, our method achieves over 30% performance improvement compared to traditional solutions like blind source separation (BSS) [44, 63, 64].

In this way, we obtain a high-quality mutual reconstruction using the residual-aided cGAN. Figure 6 illustrates a comparison between the ground truth accelerometer spectrograms when the victim pronounces the word 'mountain' with the recovered ones separated from a mixed signal. Separated data from each end undergo the same processing steps and are then recognized separately.

Algorithm 1: Iterative Underdetermined Blind Source Separation for Full-duplex Eavesdropping

Input: Mixed IMU readings A^T & G^T;
Output: Data under human-voiced sounds A^H & G^H; Data under earphone-produced sounds A^S & G^S.

Initialize $G_0^H = G^T$, Others=0, ϵ=0.01;
repeat
  Accelerometer generation: $\frac{1}{n}\sum_{n-1} G_{n-1}^H \xrightarrow{cGAN} A_n^H$;
  Spectral subtraction: $A_n^S = A^T - A_n^H$;
  Gyroscope generation: $\frac{1}{n}\sum_n A_n^S \xrightarrow{cGAN} G_n^S$;
  Updating G_n^H: $G_n^H = G^T - G_n^S$;
until $|\frac{A_n^H - A_{n-1}^H}{A_n^H}| < \epsilon \wedge |\frac{A_n^S - A_{n-1}^S}{A_n^S}| < \epsilon$;
return $A^H = A_n^H; G^H = G_n^H; A^S = A_n^S; G^S = G_n^S$.

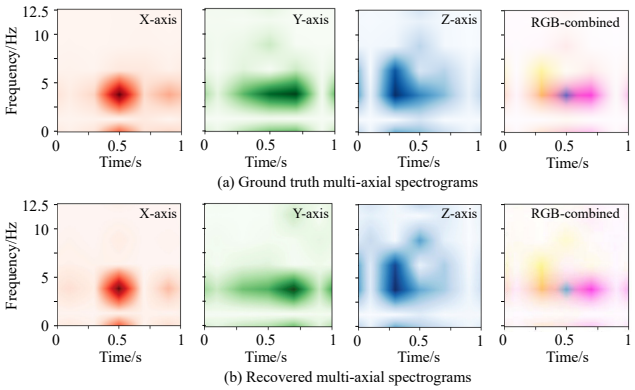


Figure 6: Comparison between reconstructed and ground truth accelerometer data in RGB spectrograms.

6 Undersampling Eavesdropping

We leverage the idea of 'space for time' by fusing multi-axis IMU data to compensate for the low sampling rate of the earphone IMU.

6.1 Multi-dimensional Data Utilization

To address the limited bandwidth caused by low sampling rates, we fully exploit the rich spatial information available in IMU signals. Unlike prior IMU-based side-channel attacks that rely on a single axis [30, 38, 39, 42] or a single sensor [30, 39, 41], we jointly leverage multi-axis data from both accelerometers and gyroscopes to compensate for the reduced temporal resolution.

6.1.1 Multi-axial Fusion via RGB Spectrogram. Previous studies either select the axis with the highest SNR [30, 39, 42] or aggregate multi-axis data into a single dimension [38, 41]. These strategies reduce computational complexity at the expense of information loss, a trade-off tolerable for eavesdropping at sampling rates above 200 Hz [30, 38, 39, 42] but detrimental to recognition accuracy when the sampling rate drops to 25 Hz. To enhance undersampling

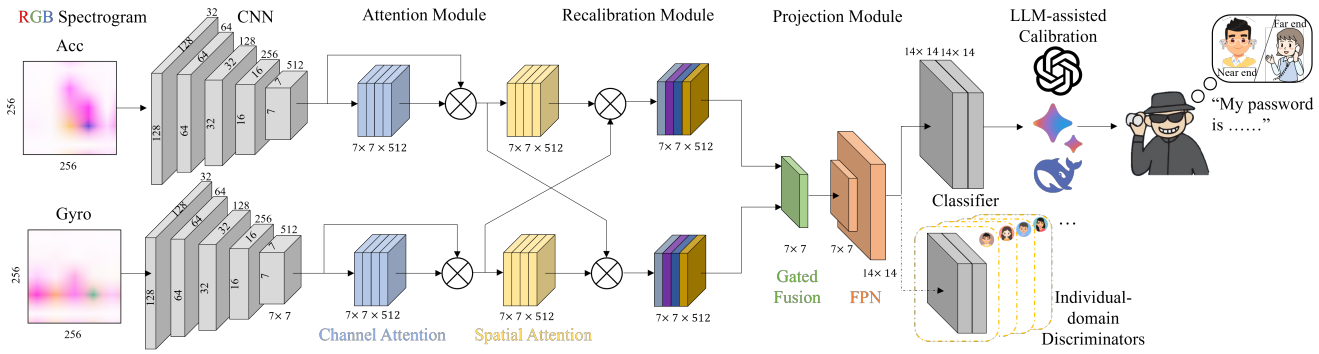


Figure 7: Architecture of our double-flow attention network, with an individual-independent adaptation module.

performance, we therefore utilize all three axes, which significantly improves information entropy (see Appendix A for proof).

Inspired by the efficient processing of three-channel RGB images in computer vision [65] we convert IMU data into three-channel RGB spectrograms instead of single-channel grayscale ones. First, we segment speech-related signals using the multiplier-based method in [38]. Then, we perform Max-Min normalization [66] and STFT on data from each axis, before combining them into an RGB image (see Figure 6 for an example with accelerometer data). The distinct spectral characteristics across different axes for the same speech input further validate the necessity of multi-axis utilization.

6.1.2 Dual-sensor Utilization. Existing IMU-based studies mainly rely on a single inertial sensor, either accelerometers or gyroscopes exclusively [25, 30, 42]. Instead, our pilot experiments in Section 3.2 demonstrate that both sensors are capable of capturing speech-induced signals. Accordingly, we integrate data from both accelerometers and gyroscopes. Specifically, RGB spectrograms for each sensor are generated following the aforementioned conversion pipeline and then fed together into the recognition network.

6.2 Attention-based Recognition Network

Figure 7 illustrates the architecture of our network to fuse accelerometer and gyroscope characteristics. Three-channel RGB spectrograms act as the inputs to our double-flow network, which utilizes a five-layer CNN as its backbone. Two attention modules [67, 68], namely the channel attention module and the spatial attention module, are recruited to determine 'what' and 'where' to attend within the feature maps. A recalibration mechanism is embedded within these attention-based modules to integrate multi-modal features from the two subnetworks. A gated fusion model [68] followed by a feature pyramid network (FPN) [69] projects features from the subnetworks into a joint space. A classifier composed of two fully-connected (FC) layers serves for IMU-to-text translation, outputting top-N candidate predictions. We detail each design component as follows.

6.2.1 Attention Module. We leverage two attention modules to improve feature representation from different dimensions. The channel attention emphasizes what to focus on, while the spatial attention determines where to focus [67, 68, 70]. The channel attention module stresses the inter-dependencies between channels. It computes the attention weight for the c-th channel in a C-channel $H \times W$-dimensional feature map $X \in \mathbb{R}^{C \times H \times W}$ as follows,

$$M_C = \sigma(W_2 \delta(W_1 \frac{\sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i,j)}{H \times W})), \tag{14}$$

where $\sigma(\cdot)$ and $\delta(\cdot)$ are Sigmoid and ReLU activation functions, and $W_i (i \in \{1, 2\})$ is the weight matrix. The spatial attention module generates a spatial attention map based on channel-wise statistics at each spatial location following

$$M_S = \sigma(Conv^{3\times3}([Pool_{avg}(X); Pool_{max}(X)])), \tag{15}$$

where $Conv^{3\times3}$ is the convolution operation with a 3×3 kernel size, $Pool_{avg}(X)$ and $Pool_{max}(X)$ represent the average pooling and max pooling results, and $[\cdot;\cdot]$ denotes channel-wise concatenation.

6.2.2 Recalibration Module. Instead of a simple parallel-branch structure that overlooks the inherent correlation, our double-flow network employs a recalibration mechanism to refine and integrate the two attentions. We recalibrate the attention weights of the two subnetworks (superscripted A and G respectively) as follows,

$$Y^A = \sigma(M_S^A \delta[Pool_{ave}(M_c^G)]), \; Y^G = \sigma(M_S^G \delta[Pool_{ave}(M_c^A)]). \tag{16}$$

6.2.3 Projection Module. To integrate features from the two subnetworks, we propose a projection module consisting of a gated fusion model [68] and a two-layer FPN [69]. In experimental comparisons with concatenation-based and co-attention-based methods [71, 72], our projection module performs better in preserving and combining multi-scale complementary information. Following this, a two-layer classifier translates the fused features to semantics.

7 Practicality Improvement

Given that adversaries often lack access to sufficiently large and diverse labeled datasets to train a generalized model, we enhance the practical generalization of our attack to ensure effectiveness across open-vocabulary and diverse individuals scenarios. We further integrate an LLM-based refinement step to resolve residual ambiguities in low-rate IMU-derived speech.

7.1 Open Vocabulary Recognition

To enable speech recovery among an open vocabulary, we propose a phoneme-based pixel-reorganization strategy.

Our approach leverages a key insight that all 48 English phonemes are present within a small corpus of common words [31, 44]. This allows a small, practical dataset to be used to learn a complete

mapping from IMU spectrograms to phoneme spectrograms. Specifically, we deconstruct the IMU spectrograms of known words into their constituent pixel columns, cataloging them by phoneme. To synthesize new word spectrograms, these columns are reassembled according to the new word's phoneme sequence. In this way, we obtain the IMU spectrograms for any target word among the open vocabulary. More details refer to Appendix B.

7.2 Individual-independent Adaptation

To transfer knowledge from our pre-trained model to a victim-specific task, we employ unsupervised domain adaptation [73]. It removes individual-dependent characteristics by just leveraging the victim's data collected during the attack phase.

We employ multiple domain discriminators to guide the representation extractor toward speaker-agnostic features. Each discriminator consists of 2 fully-connected layers, as illustrated in Figure 7. For each of the training users K paired with the target, a discriminator is trained to tell them apart. An adversarial loss [25, 74] is applied between the extractor and the discriminators. By reversing the domain loss gradient with a factor of $-\lambda$, the extractor learns to produce features that confuse the discriminators. This makes the extractor learn a shared feature space. The overall loss for the extractor is $L_f = \log \sum_{k=1}^{K} e^{L_k^s - \lambda L_k^d}$, where L_k^s and L_k^d are the classifier and domain discriminator losses for the k^{th} individual, respectively.

7.3 LLM-assisted Refinement

To correct phoneme confusion and fragmented semantics in open-vocabulary predictions, we employ a two-stage LLM calibration framework that leverages contextual reasoning. Ultimately, we yield coherent and complete sentences.

First, we feed the top-N candidate predicted by the dual-stream attention network, along with contextual clues gathered from preceding speech segments, into the LLM. The LLM performs semantic plausibility scoring. It evaluates whether each candidate conforms to the grammatical rules and the logical continuity in conversations.

Second, the LLM conducts targeted error correction for low-confidence segments marked by the inference model. For phoneme-confused words, it leverages its massive semantic corpus to distinguish homophones based on context, for instance, correcting `I'll meet you at the 'pear'` to `'pier'` when the prior conversation mentions `taking a ferry`.

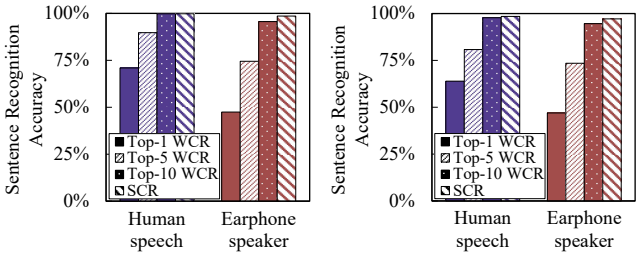
We design a lightweight prompt to focus the LLM on contextual consistency and avoid over-correction, as detailed in [75]. This prompt ensures that LLMs focus on targeted correction rather than generating irrelevant content.

8 Evaluation

8.1 Experimental Setup and Metrics

8.1.1 Setup and Dataset. Our study receives IRB approval.

Hardware. We collect 25 Hz IMU data from 10 different models of COTS earphones (7 earbuds, 3 headsets in Section 8.3.3) connected to spyware-installed host devices, including an iPhone 13 mini (iOS), a vivo X200s (AndroidOS), an iPad 9 (iPadOS 17.5.1), or a MacBook Pro 13-inch (MacOS 15.4). By default, our tests are mainly conducted on three models: Apple AirPods 3, 4, and Pro 2.



(a) Respective sentence recognition (b) Full-duplex sentence recognition

Figure 8: Sentence Eavesdropping Performance.

Data Collection. Unless otherwise specified, all experiments are conducted in a room with an ambient noise level of 44.7 dB. We recruit 52 volunteers (25 male and 27 female, aged 18~38). Specifically, three test scenarios are designed:

- Live speech test: Each volunteer wears earphones and reads (1) 30 English sentences (randomly selected from [76]) and (2) English digits (0~9) at a natural volume (approximately 66 dB on average).
- Earphone speaker test: This test includes (1) a custom-built dataset spoken by the volunteers, composed of 1000 sentences [76] in total, and (2) the AudioMNIST dataset [51] composed of English digits (0~9). Instead of following the maximum volume setup (100%, usually over 80 dB) in prior work [30, 36–43], we play two audio datasets through earphones at 61.8 dB. This volume is derived from our user habit survey (Section 8.4.1), corresponding to 50% volume for headsets or 70% for earbuds.
- Full-duplex test: Volunteers are required to speak while the audio datasets are playing through their earphones simultaneously.

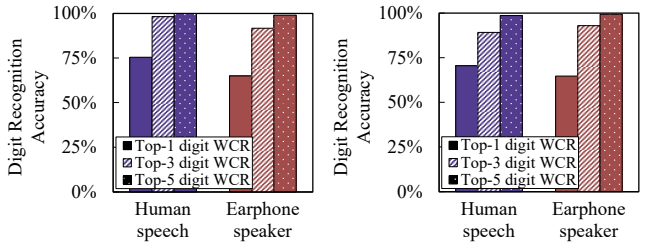
Implementation. We implement the cGAN-based reconstruction network and the attention-based recognition network using PyTorch. The networks are trained on a server equipped with an Intel(R) Core i7-14700K CPU @ 3.40 GHz and an NVIDIA GeForce RTX 4090 GPU. The collected IMU data are randomly divided into two subsets: 80% for training and 20% for testing. We adopt Gemini 2.5 Flash as the backbone model for LLM-assisted refinement.

8.1.2 Metrics. We adopt two task-specific metrics for performance evaluation: sentence correct rate (SCR) and Top-N word correct rate (WCR). To be specific, Top-N WCR quantifies the probability that the ground-truth word/digit is included in the model's top-N predicted words, and SCR measures the probability that an entire sentence is correctly recognized without any word-level errors (i.e., all words in the output match the ground-truth sentence exactly).

In addition, to assess our signal separation methods (Section 5.2), we employ two image similarity metrics that quantify discrepancies between separated spectrograms and the ground truth. Structural Similarity Index Measure (SSIM) [77] evaluates structural quality by comparing luminance, contrast, and structure. Learned Perceptual Image Patch Similarity (LPIPS) assesses perceptual differences using deep models that mimic human vision. High similarity is indicated by SSIM > 95% or LPIPS < 0.2 (virtually imperceptible), while scores of SSIM > 90% or LPIPS < 0.4 suggest only slight blurring.

8.2 Overall Performance

8.2.1 Sentence Recognition. We first evaluate sentence recognition performance with results shown in Figure 8.



(a) Respective Digit recognition (b) Full-duplex digit recognition

Figure 9: Digit Eavesdropping Performance.

On Live Human Speech: Our attack achieves strong performance in open-vocabulary speech recognition via human-to-IMU eavesdropping, with a 70.9% top-1 WCR. In particular, the top-10 WCR approaches 100%, indicating the correct sentence is nearly always contained within the top-10 prediction. Without our refinement, the proportion of complete sentences where every single word appears within the raw Top-1, Top-5, and Top-10 candidate pools is 9.8%, 45.4%, and 90.4%, respectively. Leveraging these top-10 candidates, our LLM-assisted refinement framework identifies the correct sentence with a high SCR of 99.7%.

On Earphone Speakers: For speech from earphone speakers, we attain 49.5% top-1 WCR , attributed to the lower SNR (see Table 1). Nevertheless, the top-10 WCR exceeds 95%. Even though individual words may be omitted from the top-10 candidates, our LLM-assisted refinement framework identifies the correct sentence via contextual correction, achieving a high SCR of 97.1%.

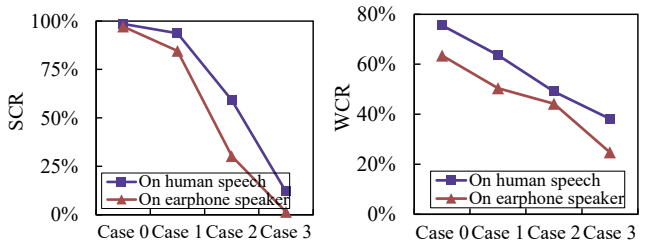
Full-duplex Performance: Under the same setup, we extend our evaluation to full-duplex eavesdropping, following an assessment on reconstructed and separated spectrograms.

We establish a cGAN in Section 5.2.2 to mutually reconstruct IMU data. The model achieves a high SSIM of up to 96.70%±1.73% and a low LPIPS of merely 0.17±0.06. The results indicate a superior generative capability of our network. We next evaluate the spectrogram separation quality of the iterative algorithm presented in Section 5.2.3. The method yields an SSIM of 91.57% ±1.83% and an LPIPS of 0.20±0.04, which are acceptable for further recognition.

As shown in Figure 8(b), we evaluate word recognition performance on the separated signals. For live human speech, the WCR loss is within 7%, while that for earphone speaker eavesdropping is within 1%. For SCR, the full-duplex accuracy loss is below 2%. This validates the feasibility of underdetermined source separation using only earphone IMUs for full-duplex eavesdropping.

8.2.2 Digit Recognition. Digits (usually used as passwords) lack contextual relationships, leaving our LLM-assisted refinement framework unable to enhance their recognition. We thus evaluate digit recognition performance separately, with results in Figure 9.

Each Stream Performance: Figure 9(a) shows the top-N accuracy to recognize digits from each source, respectively. The top-1 accuracy of human-voiced digits reaches 75%, top-3 hits 98%, and top-5 achieves 100%. For speaker-to-IMU side channels, we attain 62.9% top-1 accuracy, with top-5 accuracy over 99%. These results demonstrate that our side-channel attack significantly reduces the k-digit password search space from 10^k to below 5^k, posing a considerable security threat. A detailed comparison of the search space and entropy in digit passwords see Appendix C.



(a) Sentence inference (b) Digit recognition

Figure 10: Ablation Study.

Full-duplex Performance: As illustrated in Figure 9(b), we perform digit recognition on the signals separated via our iterative scheme. For live human speech recognition, the additional accuracy loss is within 5%. For earphone speaker eavesdropping, the top-1 accuracy remains at 64.7%, which exhibits little performance degradation (of merely 0.2%).

8.2.3 Ablation Study. We conduct an ablation experiment on our improvements to the cGAN and recognition networks.

We add a residual connection to our cGAN's U-Net to mitigate gradient vanishing issues. Without this modification, a traditional U-Net produces low-quality spectrograms with 75.05%±0.31% SSIM and 0.38±0.06 LPIPS experimentally.

Figure 10 validates the individual contributions of our three specialized modules in the recognition network. Starting with the complete network as a baseline (Case 0), we sequentially removed the recalibration module (Case 1), then the fusion projection module (Case 2), and finally weakened the attention module (Case 3) by using single-axis (grayscale) instead of multi-axial (RGB) spectrograms. In the task of sentence inference, the SCR decreases consistently. This shows that without accurate candidates provided by our attention model, adversaries, even by the aid of LLMs, cannot achieve effective eavesdropping independently. In Figure 10(b), the consistent decline in digit recognition accuracy further confirms the effectiveness and necessity of each component in our framework.

8.3 Scalability Study

To demonstrate real-world scalability and robustness, we focus our subsequent evaluations on digit recognition rather than sentence recognition, unless otherwise specified.

8.3.1 Open-Vocabulary Performance. We recruit four participants to read 60 randomly-selected new sentences, comprising over 200 new words not present in the previous training setup. As shown in Figure 14(a), our model achieves an average top-1 WCR of 68.8% for live human-to-IMU eavesdropping and 47.8% for earphone speaker-to-IMU eavesdropping on these unseen words. Additionally, the top-10 WCR exceeds 97%. The overall loss in accuracy is below 3% among unseen words. Note that these top-N WCR values are calculated exclusively for the new words, with no overlap with previously seen vocabulary. Even for sentences containing unseen words, we achieve high SCRs of over 98%. These results demonstrate the open-vocabulary capability of our attack.

8.3.2 Across-Individual Generalization. The experiment is extended to include four new participants (two male and two female), whose data are absent from the pre-trained model's training set. As shown

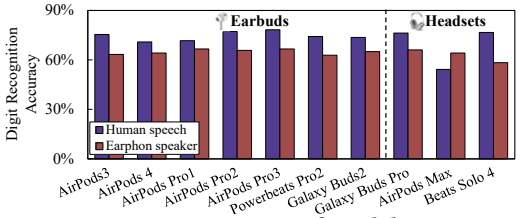


Figure 11: Impact of models.

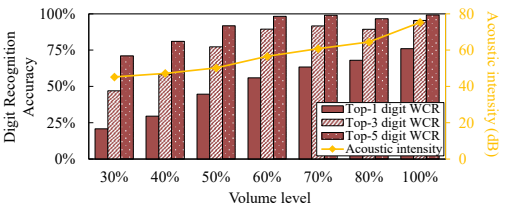


Figure 12: Impact of media volume setting.

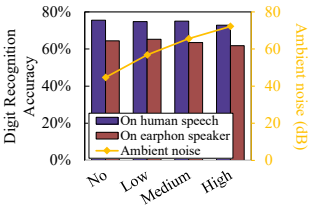
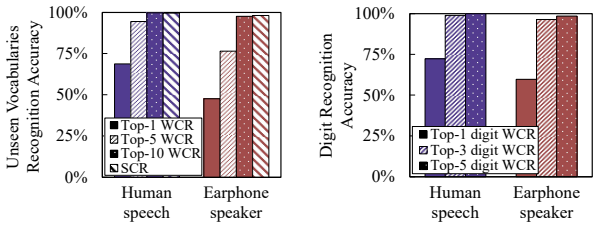


Figure 13: Ambient Impact.



(a) Among unseen vocabularies (b) Across new individuals

Figure 14: Scalability Performance.

in Figure 14(b), our attack yields recognition accuracies of 72.8% for digits in live speech eavesdropping, and 59.8% for digits in earphone speaker eavesdropping, with top-N digit WCR of over 98%. This performance is comparable to that of previously-seen individuals.

8.3.3 Various-Earphone-Model Performance. Our attack generalizes across various models, including both earbuds and headsets. Figure 11 presents the performance in digit recognition among 10 different models, covering popular brands like Apple AirPods, Samsung Galaxy, and Beats.

On average, across six tested earbuds, we achieve average accuracies of 74.6% (live speech) and 64.9% (speaker audio). The three headsets average 68.6% and 60.4%, respectively. The higher vulnerability of earbuds stems from their compact design, where the IMU is closer to the speaker with higher signal leakage, and is more affected by facial movements. Despite these differences, major earphones are vulnerable to our full-duplex attack.

In short, we realize individual-independent eavesdropping on various earphone models across open vocabularies.

8.4 Robustness Study

8.4.1 Impact of Media Volume Setting. The volume level setting directly decides the acoustic intensity and the resultant SNR in the speaker-to-IMU side channel. Following a survey on users' preferred volumes to determine the defaults in our experiments, we evaluate our attacks across various volume levels.

To determine a realistic volume level for our experiments, we deviate from previous studies that default to maximum volume [30, 38–42], a setting rarely used in daily life due to comfort and hearing health concerns. We survey 201 participants (100 males, 101 females, aged 18~49) to investigate their typical earphone usage in the last year. Based on the statistics from the Apple `Health` application [78], the average listening exposure is 61.8 dB. Accordingly, we calibrate our experimental volume to approximate this level. We set earbuds (e.g., AirPods) to 70% of maximum volume and headsets (AirPods Max) to 50%, whose intensity measures around 61.5 dB.

As shown in Figure 12, our approach correctly recognizes more digits as the volume increases on Apple earbuds. At the maximum volume level, the top-1 accuracy peaks at 76.0%, approaching the SOTA work [38] using 200 Hz. At 50% volume, our method maintains a top-1 accuracy exceeding 40% and a top-5 accuracy of 92.8%. Even down to 30% volume, the top-5 accuracy remains above 71.0%, although the top-1 accuracy drops to 18.8%. However, the low-volume setting measures only 45.3 dB, which is slightly higher than the ambient noise level. This setting is quite scarce in daily earphone usage due to the insufficient SNR for human ears. In short, our attacks demonstrate applicability across real-world settings.

8.4.2 Impact of Users' Movements. To assess robustness to real-world motion, we test our attack on four volunteers (two males and two females) who are performing large-scale body (walking, running, bicycling, and driving) and small-scale head (nodding, shaking, chewing, and yawning) motion. Our proposed motion artifact elimination technique proves highly effective. Experimental results demonstrate that live human speech eavesdropping accuracy was maintained at over 74% during various movement, while earphone speaker eavesdropping accuracy remained above 60% in spite of various motion artifacts.

8.4.3 Impact of Environment. We repeat the attacks when victims are located in four environments with different levels of ambient noise: (1) a quiet room (no-noise, 44.7 dB), (2) a lab with people talking (low-noise, 56.8 dB), (3) a restaurant with music playback (medium-noise, 65.6 dB), and (4) a crowded Metro station (high-noise, 72.2 dB). Four volunteers (two males and two females, aged from 20 to 25) are asked to wear Apple AirPods 3 and speak digits at their natural voice volumes. Besides, the earphones play AudioMNIST audio [51]. As illustrated in Figure 13, ambient noise poses little effect on the duplex eavesdropping performance on both live human speech and earphone speakers.

8.5 End-to-end Attack Case Study

We conduct a case study on phone-based customer service scenarios. Four pairs of participants randomly perform 240 typical conversations, covering common requests like bill dispute resolution (as exemplified in [75]), plan inquiries, and account management. Our attack achieves 100% accuracy in recognizing these continuous dialogues. This enables adversaries to reconstruct phone interaction flows. Leveraging captured conversational context, attackers can identify user intentions and activities.

Furthermore, our attack enables adversaries to infer high-risk sensitive information, such as account numbers and verification codes. Random 8-digit passwords are embedded into both ends of the 240 tested dialogues. As shown in Table 2, our attack infers multiple correct digits. This capability drastically reduces the cryptographic key space. For an 8-digit PIN, the original 10^8 possible combinations narrow to $\sim 3^8$ (>99.99% reduction) for live speech and

Table 2: Number of correctly inferred digits

Number	0	1	2	3	4	5	6	7	8
Live speech	0	2	2	9	21	52	72	60	22
Earphone speaker	0	1	8	24	70	61	53	18	5

$\sim 5^8$ (>99.6% reduction) for earphone playback. Correspondingly, the password entropy declines from 26.6 bits to only 9.27 bits for human-voiced passwords and to 12.63 bits for earphone-produced passwords. Detailed calculations see Appendix C. This directly demonstrates the attack's practical, non-trivial threat, in which it lowers the barrier for brute-force or targeted guessing attacks.

8.6 Comparison with SOTA Attacks

Compared to existing side-channel attacks on mobile devices, our work is the first to achieve full-duplex eavesdropping. This attack relies solely on built-in IMUs in COTS earphones, without requiring victim cooperation in terms of permissions or device placement. Beyond these inherent advantages, we further conduct a comprehensive performance comparison with SOTA side-channel attacks.

8.6.1 With Human-to-IMU Attacks. Face-Mic [25] recognizes live human speech via AR/VR devices sampled at 1000 Hz. It achieves 96% accuracy among 11 digits and 10 words. EarSpy [26] performs eavesdropping using self-made earphone prototypes rather than COTS. Among a vocabulary of 120 words, EarSpy achieves 92.5% accuracy at a 1000 Hz sampling rate, and 85.1% accuracy at 400 Hz. In addition, we implement the attacking methods in Face-Mic [25] and EarSpy [26] on COTS earphones sampled at 25 Hz among the identical digit vocabulary, and their accuracies drop to below 60%. As compared in Table 3, our approach exploits built-in IMUs of COTS earphones operating at 25 Hz, less than one-tenth of the sampling rate used in the above two attacks. Despite this significant hardware limitation, our system achieves an acceptable accuracy of over 70% across the open vocabulary.

8.6.2 With Speaker-to-IMU Attacks. We compare the digit recognition accuracy of our method with SOTA side-channel attacks [30, 38, 41] in Table 4, where all experiments are conducted on the identical AudioMNIST dataset to ensure fair comparison.

Our approach exhibits superior efficiency. At an extremely low sampling rate of merely 25 Hz, we achieve an accuracy exceeding 60%, outperforming SOTA methods that only reach 30%∼55% accuracy even when operating at a higher sampling rate of 50 Hz [30, 38, 41]. Moreover, this performance of up to 75% is highly competitive with SOTA attacks [30, 38], which report 78% accuracy but demand substantially higher sampling rates (200∼500 Hz). Even when under-sampled to 5 Hz, an extreme low-rate scenario, our attack still surpasses the SOTA benchmark proposed in [41]. In contrast to SOTA methods that require smartphones to operate at maximum volume (>80 dB), our attack is implemented on earphones with a mere 61.5 dB output. This results in a significantly lower side-channel SNR, yet our method maintains superior performance. Such robustness under constrained conditions is attributed to our proposed multi-dimensional signal utilization.

In summary, our attack achieves performance marginally lower than existing simplex methods, despite using a sampling rate at least eight times lower (25 Hz vs. >200 Hz). Moreover, our approach offers unique advantages of full-duplex eavesdropping capabilities across an open vocabulary.

Table 3: Human-to-IMU eavesdropping comparison

	Victim	Sensor		Vocabulary	Accuracy
[25]	AR/VR	Acc+ Gyro	1000 Hz	21 words	96%
			200 Hz		93%
			25 Hz	10 digits*	57.5%
[26]	Earbuds prototypes	Acc	1000 Hz	120 words	92.5%
			400 Hz		85.1%
			25 Hz	10 digits*	31.5%
Our	COTS earphones	Acc+ Gyro	25 Hz	10 digits	75.5%
				open	70.9%

*: We reimplement the attacking methods in [25] and [26] on COTS earphones sampled at 25 Hz.

Table 4: Speaker-to-IMU 10-digit eavesdropping comparison

	Victim	Sensor	Volume	Fs	Accuracy
[30]	Smartphone	Acc	100%	50 Hz	30%
				500 Hz	78%
[38]	Smartphone	Acc+ Gyro	100%	200 Hz	78.8%
				50 Hz	54.7%
				20 Hz	26.9%
[41]	Smartphone	Acc	100%	200 Hz	77.8%
				50 Hz	56.7%
				5 Hz	32.7%
Our	Earphone	Acc+ Gyro	100%	25 Hz	76.0%
			50%	25 Hz	62.9%
				5 Hz	34.4%

9 Countermeasure and Discussion

Potential Countermeasure. To defend against IMU-based eavesdropping without degrading user experience, we propose an ultrasonic jamming strategy. While direct application of audible noise (e.g., background music [41]) proves effective, such methods inevitably compromise user experience. Instead, our approach leverages the IMU's sensitivity to ultrasound [45, 79, 80] to reduce the speech SNR silently. This provides an effective defense that balances security and user experience.

Future Improvements. There is potential to reconstruct high-frequency speech signals from 25 Hz IMU data using encoder-decoder architecture [30], cGAN [42], and super-resolution mechanism [40]. Nevertheless, there is much room to be desired in recovering audio from the narrow 12.5 Hz band of IMU data. In addition, we will explore IMU side-channels to infer other sensitive information, such as user identity [25, 30, 38] and location [39, 41].

10 Conclusion

We realize the first full-duplex eavesdropping attack on COTS earphones, exploiting zero-permission IMU sensors. By leveraging the correlation between the accelerometer and gyroscope, our method simultaneously eavesdrops on both participants in bidirectional earphone activities. Operating at a low 25 Hz sampling rate, the attack executes stealthily in the background without being detected. We also propose a countermeasure to mitigate this threat.

Acknowledgements

This paper is partially supported by National Natural Science Foundation of China (Grant No. 62502235), National Science Fund for Distinguished Young Scholars of China (Grant No. 62125203), Natural Science Foundation of Jiangsu Province of China (Grant No. BK20240615). This paper was edited for grammar using Gemini 3.

Open Science

The source code is publicly available at https://anonymous.4open.science/r/Turning-Everyday-Earphones-into-a-Full-Duplex-Speech-Eavesdropping-Platform-via-Zero-permission-IMU-EZ85.

To protect data privacy and meet ethical requirements, the full dataset is not publicly released. A representative subset of the data is provided as examples to demonstrate the artifact's functionality.

Ethical Considerations

This study follows the approved Institutional Review Board (IRB) protocol and relevant ethical guidelines. To the best of our knowledge, it raises no ethical concerns.

This paper investigates a novel side-channel risk from IMUs in COTS earphones. We demonstrate that both live human speech and earphone audio playback can be inferred without leveraging the microphone interface. On contemporary platforms, this capability often does not require explicit user authorization. It is important to emphasize that our goal is not to facilitate covert eavesdropping. Instead, we aim to characterize a side-channel that remains underestimated in prevailing sensor-access frameworks and threat models. Our results indicate that users face potential privacy leakage even when they deny microphone access, and that such leakage may indirectly affect their remote communication partners.

All experiments conducted in this study, i.e., sensor data collection from human-worn headphones, have been reviewed and approved by the Medical Ethics Committee of our institution. The research strictly adheres to the approved IRB protocol and national regulations governing biomedical research involving human subjects. In conducting this research, we obtain informed consent from all participants. Data collection is limited to what is strictly necessary for feasibility evaluation, and all collected data are anonymized. Participants retain the right to access their personal data and to request its deletion at any time. No data is made publicly available without explicit additional consent. Given the potential for misuse, we do not release the datasets. Instead, following open science principles, we focus on characterizing the technical boundaries and system-level causes of the leakage using openly available artifacts.

To address the identified risk, we propose several defensive strategies that can be implemented across hardware, operating systems, and applications. Future work will refine these protections and examine whether similar vulnerabilities exist in other devices.

We have reported this vulnerability to affected vendors, including Apple, Beats, and Samsung, and shared the potential countermeasures detailed in our paper. We call for greater attention from both manufacturers and the public to side-channel threats in commercial audio devices.

Overall, this work advances the privacy-aware design of wearable and mobile systems by illuminating and addressing a previously underappreciated attack vector.

References

[1] iOShacker. How to enable or disable microphone on iphone. https://ioshacker.com/how-to/enable-disable-microphone-iphone, 2025.

[2] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. Backdoor: Making microphones hear inaudible sounds. In *Proceedings of ACM MobiSys*, pages 2–14, 2017.

[3] Ming Gao, Yike Chen, Yajie Liu, Jie Xiong, Jinsong Han, and Kui Ren. Cancelling speech signals for speech privacy protection against microphone eavesdropping. In *Proceedings of ACM MobiCom*, pages 36:1–36:16, 2023.

[4] Ke Sun, Chen Chen, and Xinyu Zhang. "Alexa, stop spying on me!": Speech privacy protection against voice assistants. In *Proceedings of ACM SenSys*, pages 298–311, 2020.

[5] Lingkun Li, Manni Liu, Yuguang Yao, Fan Dang, Zhichao Cao, and Yunhao Liu. Patronus: Preventing unauthorized speech recordings with support for selective unscrambling. In *Proceedings of ACM SenSys*, pages 245–257, 2020.

[6] Shengyu Li, Mengchen Teng, Boyu Li, Songfan Li, Xiandong Shao, Chong Zhang, and Li Lu. Hedgehog: Pushing the range limits of ultrasonic microphone jammers. In *Proceedings of ACM MobiCom*, 2025.

[7] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, Kun Wang, and Wenyao Xu. Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface. In *Proceedings of ACM MobiSys*, page 14–26, 2019.

[8] Suryoday Basak and Mahanth Gowda. mmspy: Spying phone calls using mmwave radars. In *Proceedings of IEEE SP*, pages 1211–1228, 2022.

[9] Chao Wang, Feng Lin, Hao Yan, Tong Wu, Wenyao Xu, and Kui Ren. Vibspeech: exploring practical wideband eavesdropping via bandlimited signal of vibration-based side channel. In *Proceedings of USENIX Security*, 2024.

[10] Pengfei Hu, Wenhao Li, Riccardo Spolaor, and Xiuzhen Cheng. mmecho: A mmwave-based acoustic eavesdropping method. In *Proceedings of IEEE SP*, pages 1840–1856, 2023.

[11] Suryoday Basak, Abhijeeth Padarthi, and Mahanth Gowda. mmwave-whisper: Phone call eavesdropping and transcription using millimeter-wave radar. In *Proceedings of IEEE ICASSP*, pages 1–5, 2025.

[12] Liang Wang, Aoyang Chang, Boqun Liu, Zhiwen Yu, Pengfei Hu, Jia Liu, Yao Zhang, and Bin Guo. mmhse: Enhanced eavesdropping attack on headsets leveraging COTS mmwave radar. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 9(2):51:1–51:27, 2025.

[13] Sriram Sami, Sean Rui Xiang Tan, Yimin Dai, Nirupam Roy, and Jun Han. Lidarphone: acoustic eavesdropping using a lidar sensor: poster abstract. In *Proceedings of ACM SenSys*, pages 701–702, 2020.

[14] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J. Mysore, Frédo Durand, and William T. Freeman. The visual microphone: passive recovery of sound from video. *ACM Trans. Graph.*, 33(4), 2014.

[15] Yunzhong Chen, Jiadi Yu, Linghe Kong, Hao Kong, Yanmin Zhu, and Yi-Chao Chen. Rf-mic: Live voice eavesdropping via capturing subtle facial speech dynamics leveraging RFID. *Proceedings of IMWUT/Ubicomp*, 7(2):49:1–49:25, 2023.

[16] Teng Wei, Shu Wang, Anfu Zhou, and Xinyu Zhang. Acoustic eavesdropping through wireless vibrometry. In *Proceedings of ACM MobiCom*, page 130–141, 2015.

[17] Chuyu Wang, Lei Xie, Yuancan Lin, Wei Wang, Yingying Chen, Yanling Bu, Kai Zhang, and Sanglu Lu. Thru-the-wall eavesdropping on loudspeakers via RFID by capturing sub-mm level vibration. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(4):182:1–182:25.

[18] Ziqi Wang, Zhe Chen, Akash Deep Singh, Luis Garcia, Jun Luo, and Mani B. Srivastava. Uwhear: through-wall extraction and separation of audio vibrations using wireless signals. In *Proceedings of ACM SenSys*, page 1–14, 2020.

[19] Yu Rong, Sharanya Srinivas, Adarsh A. Venkataramani, and Daniel W. Bliss. UWB radar vibrometry : An RF microphone. In *Proceedings of IEEE ACSCC*, pages 1066–1070, 2019.

[20] IDC Corporate. Worldwide quarterly wearable device tracker. https://www.idc.com/getdoc.jsp?containerId=IDC_P31315, 2025.

[21] Apple Inc. Cmheadphonemotionmanager - an object that starts and manages headphone motion services. https://developer.apple.com/documentation/coremotion/cmheadphonemotionmanager, 2024.

[22] James Archer. 360 audio on samsung galaxy buds pro: How it works and how to use it. https://www.tomsguide.com/how-to/360-audio-on-samsung-galaxy-buds-pro-how-it-works-and-how-to-use-it, 2021.

[23] Apple Inc. Control spatial audio and head tracking. https://support.apple.com/guide/airpods/control-spatial-audio-and-head-tracking-dev00eb7e0a3/web, 2020.

[24] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. Imuposer: Full-body pose estimation using imus in phones, watches, and earbuds. In *Proceedings of ACM CHI*, pages 529:1–529:12, 2023.

[25] Cong Shi, Xiangyu Xu, Tianfang Zhang, Payton Walker, Yi Wu, Jian Liu, Nitesh Saxena, Yingying Chen, and Jiadi Yu. Face-mic: inferring live speech and speaker identity via subtle facial dynamics captured by AR/VR motion sensors. In *Proceedings of ACM MobiCom*, pages 478–490, 2021.

[26] Yetong Cao, Fan Li, Huijie Chen, Xiaochen Liu, Chunhui Duan, and Yu Wang. I can hear you without a microphone: Live speech eavesdropping from earphone motion sensors. In *Proceedings of IEEE INFOCOM*, pages 1–10, 2023.

[27] Lan Sun, Songpengcheng Xia, Junyuan Deng, Jiarui Yang, Zengyuan Lai, Qi Wu, and Ling Pei. Suite-in: Aggregating motion features from apple suite for robust inertial navigation. In *Proceedings of IEEE ICRA*, pages 3625–3631, 2025.

[28] Stack Overflow. What frequency standard does cmheadphonemotionmanager's startdevicemotionupdates() has? https://stackoverflow.com/questions/71612270/what-frequency-standard-does-cmheadphonemotionmanagers-startdevicemotionupdates, 2025.

[29] Shengzhe Lyu, Yongliang Chen, Di Duan, Renqi Jia, and Weitao Xu. Earda: Towards accurate and data-efficient earable activity sensing. In *Proceedings of IEEE CSCAIoT*, pages 1–7, 2024.

[30] Zhongjie Ba, Tianhang Zheng, Xinyu Zhang, Zhan Qin, Baochun Li, Xue Liu, and Kaili Ren. Learning-based practical smartphone eavesdropping with built-in accelerometer. In *Proceedings of NDSS*, 2020.

[31] Antonie Cohen. *The phonemes of English*. Springer, 1965.

[32] Snoy Inc. Speak-to-chat. https://www.sony.jp/feature/owner/hp/fy22_howto_speak-to-chat/?linkId=100000164635431, 2021.

[33] Mike Kobrin. Headphone conversation awareness mode: How it works and why you need it. https://www.cnet.com/tech/mobile/headphone-conversation-awareness-mode-how-it-works-and-why-you-need-it/, 2025.

[34] Weigao Su, Daibo Liu, Taiyuan Zhang, and Hongbo Jiang. Towards device independent eavesdropping on telephone conversations with built-in accelerometer. *Proceedings of IMWUT/Ubicomp*, 5(4):177:1–177:29, 2021.

[35] Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. Unvoiced: Designing an llm-assisted unvoiced user interface using earables. In *Proceedings of ACM SenSys*, page 784–798, 2024.

[36] S. Abhishek Anand and Nitesh Saxena. Speechless: Analyzing the threat to speech privacy from smartphone motion sensors. In *Proceedings of IEEE S&P*, pages 1000–1017, 2018.

[37] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. Gyrophone: Recognizing speech from gyroscope signals. In *Proceedings of USENIX Security*, pages 1053–1067, 2014.

[38] Ming Gao, Yajie Liu, Yike Chen, Yimin Li, Zhongjie Ba, Xian Xu, and Jinsong Han. Inertiear: Automatic and device-independent imu-based eavesdropping on smartphones. In *Proceedings of IEEE INFOCOM*, pages 1129–1138, 2022.

[39] Ke Sun, Chunyu Xia, Songlin Xu, and Xinyu Zhang. Stealthyimu: Extracting permission-protected private information from smartphone voice assistant using zero-permission sensors. In *Proceedings of NDSS*, 2023.

[40] Lei Wang, Meng Chen, Li Lu, Zhongjie Ba, Feng Lin, and Kui Ren. Voicelistener: A training-free and universal eavesdropping attack on built-in speakers of mobile devices. *Proceedings of IMWUT/Ubicomp*, 7(1):32:1–32:22, 2023.

[41] Qingsong Yao, Yuming Liu, Xiongjia Sun, Xuewen Dong, Xiaoyu Ji, and Jianfeng Ma. Watch the rhythm: Breaking privacy with accelerometer at the extremely-low sampling rate of 5hz. In *Proceedings of ACM CCS*, pages 1776–1790, 2024.

[42] Pengfei Hu, Hui Zhuang, Panneer Selvam Santhalingam, Riccardo Spolaor, Parth Pathak, Guoming Zhang, and Xiuzhen Cheng. Accear: Accelerometer acoustic eavesdropping with unconstrained vocabulary. In *Proceedings of IEEE SP*, pages 1757–1773, 2022.

[43] Ahmed Najeeb, Abdul Rafay, Muhammad Hamad Alizai, and Naveed Anwar Bhatti. Glitch in time: Exploiting temporal misalignment of IMU for eavesdropping. In *Proceedings of ACM ASIA CCS*, pages 530–541, 2025.

[44] Yunzhong Chen, Jiadi Yu, Yingying Chen, Linghe Kong, Yanmin Zhu, and Yi-Chao Chen. Rfspy: Eavesdropping on online conversations with out-of-vocabulary words by sensing metal coil vibration of headsets leveraging RFID. In *Proceedings of ACM MOBISYS*, pages 169–182, 2024.

[45] Timothy Trippel, Ofir Weisse, Wenyuan Xu, Peter Honeyman, and Kevin Fu. WALNUT: waging doubt on the integrity of MEMS accelerometers with acoustic injection attacks. In *Proceedings of IEEE European SP*, 2017.

[46] R.G. Vaughan, N.L. Scott, and D.R. White. The theory of bandpass sampling. *IEEE Transactions on Signal Processing*, 39(9):1973–1984, 1991.

[47] Prabhavathi C N, Gagan G Kashyap, and Gagana N. Compressive sensing and its application to speech signal processing. In *Proceedings of NMITCON*, 2023.

[48] Cameron L Jones, Ajit Achuthan, and Byron D Erath. Modal response of a computational vocal fold model with a substrate layer of adipose tissue. *Journal of the Acoustical Society of America*, 137(2):58–64, 2015.

[49] Xingzhe Song, Hongshuai Li, and Wei Gao. Myomonitor: Evaluating muscle fatigue with commodity smartphones. *Smart Health*, 19:100175, 2021.

[50] Estefania Munoz Diaz, Oliver Heirich, Mohammed Khider, and Patrick Robertson. Optimal sampling frequency and bias error modeling for foot-mounted imus. In *Proceedings of International Conference on Indoor Positioning and Indoor Navigation*, pages 1–9, 2013.

[51] Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Interpreting and explaining deep neural networks for classification of audio signals. *CoRR*, abs/1807.03418, 2018.

[52] S Abhishek Anand, Chen Wang, Jian Liu, Nitesh Saxena, and Yingying Chen. Spearphone: A lightweight speech privacy exploit via accelerometer-sensed reverberations from smartphone loudspeakers. In *Proceedings of ACM WiSec*, pages 288–299, 2021.

[53] Ahmed El Hady and Benjamin B Machta. Mechanical surface waves accompany action potential propagation. *Nature communications*, 6(1):6697, 2015.

[54] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *Computer Science*, page 2672–2680, 2014.

[55] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of NeurIPS*, 2014.

[56] Ke Sun and Xinyu Zhang. Ultrase: single-channel speech enhancement using ultrasound. In *Proceedings of ACM MobiCom*, 2021.

[57] Guoming Zhang, Zhijie Xiang, Heqiang Fu, Yanni Yang, and Pengfei Hu. Echolight: Sound eavesdropping based on ambient light reflection. In *Proceedings of IEEE INFOCOM*, 2024.

[58] Shanshan Yu, Ju Liu, Jing Wang, and Inam Ullah. Adaptive double-threshold cooperative spectrum sensing algorithm based on history energy detection. *Wireless Communication and Mobile Computing*, 2020:4794136:1–4794136:12, 2020.

[59] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Proceedings of NeurIPS*, 2017.

[60] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of IEEE/CVF CVPR*, 2017.

[61] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE/CVF CVPR*, 2016.

[62] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120, 1979.

[63] Ming Gao, Feng Lin, Weiye Xu, Muertikepu Nuermaimaiti, Jinsong Han, Wenyao Xu, and Kui Ren. Deaf-aid: mobile iot communication exploiting stealthy speaker-to-gyroscope channel. In *Proceedings of ACM MobiCom*, 2020.

[64] Lei Wang, Xingwei Wang, Yu Zhang, Xiaolei Ma, Haipeng Dai, Yong Zhang, Zhijun Li, and Tao Gu. Accurate blood pressure measurement using smartphone's built-in accelerometer. *Proceedings of IMWUT/Ubicomp*, 8(2):1–28, 2024.

[65] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of IEEE/CVF CVPR*, 2017.

[66] Dorian Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann Publishers Inc., 1999.

[67] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of IEEE/CVF CVPR*, 2018.

[68] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. In *Proceedings of ECCV*, 2018.

[69] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of IEEE/CVF CVPR*, 2017.

[70] Tiantian Liu, Feng Lin, Chao Wang, Chenhan Xu, Xiaoyu Zhang, Zhengxiong Li, Wenyao Xu, Ming-Chun Huang, and Kui Ren. Wavoid: Robust and secure multi-modal user identification via mmwave-voice mechanism. In *Proceedings of ACM UIST*, 2023.

[71] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Proceedings of NeurIPS*, 2016.

[72] Tiantian Liu, Ming Gao, Feng Lin, Chao Wang, Zhongjie Ba, Jinsong Han, Wenyao Xu, and Kui Ren. Wavoice: A noise-resistant multi-modal speech recognition system fusing mmwave and audio signals. In *Proceedings of ACM SenSys*, 2021.

[73] Han Zhao, Shanghang Zhang, Guanhang Wu, José M. F. Moura, João Paulo Costeira, and Geoffrey J. Gordon. Adversarial multiple source domain adaptation. In *Proceedings of NeurIPS*, 2018.

[74] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[75] Anonymous Github. Turning-everyday-earphones-into-a-full-duplex-speech-eavesdropping-platform-via-zero-permission-imu. https://anonymous.4open.science/r/Turning-Everyday-Earphones-into-a-Full-Duplex-Speech-Eavesdropping-Platform-via-Zero-permission-IMU-EZ85/README.md, 2026.

[76] SEC. 1000 english sentences used in daily life. https://speakenglishcenter.com/english-sentences-used-in-daily-life/, 2023.

[77] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[78] Apple Support. Check your headphone audio levels on iphone. https://support.apple.com/en-gb/guide/iphone/iph0596a9152/ios, 2024.

[79] Yunmok Son, Hocheol Shin, Dongkwan Kim, Young-Seok Park, Juhwan Noh, Kibum Choi, Jungwoo Choi, and Yongdae Kim. Rocking drones with intentional sound noise on gyroscopic sensors. In *Proceedings of USENIX Security*, 2015.

[80] Yazhou Tu, Zhiqiang Lin, Insup Lee, and Xiali Hei. Injected and delivered: Fabricating implicit control over actuation systems by spoofing inertial sensors. In *Proceedings of USENIX Security*, 2018.

A Complementarity of Multi-axis Signals

Fusing all six IMU axes (3-axis accelerometer/gyroscope) compensates for undersampling-induced information loss without overhead.

Proof. The 25 Hz sampling rate bounds each single-axis signal to $0 \sim 12.5$ Hz, yielding sparse and incomplete articulatory characterization. Consequently, the Shannon entropy $H(X_i)$ of any single axis is bounded above: $H(X_i) \leq \xi$.

These six axes exhibit structured interdependence, neither independent nor perfectly correlated. The joint entropy expands as:

$$H(X_1, \ldots, X_6) = \sum_{i=1}^{6} H(X_i) - \sum_{i<j} I(X_i; X_j) + \sum_{i<j<k} I(X_i; X_j; X_k) - \cdots, \tag{17}$$

where $I(\cdot)$ denotes mutual information. For any two axes,

$$0 < I(X_i; X_j) < H(X_i), \tag{18}$$

indicating partial correlation: they share common source information while retaining unique complementary components. Higher-order terms decay rapidly. Hence,

$$H(X_1, \ldots, X_6) > \max_i H(X_i). \tag{19}$$

Thus, the fused multi-axis representation strictly contains more information than any single channel. □

This complementary information compensates for per-channel undersampling loss and provides sufficient entropy for accurate open-vocabulary recognition.

B Phoneme-Based Pixel Reorganization Enables Open-Vocabulary Recognition

For any open-vocabulary word $w \in V$ (where V denotes the set of all English words), its IMU spectrogram can be constructed by reorganizing pre-learned phoneme-to-pixel mappings. This process leverages the insight that the set of 48 English phonemes is fully covered within a compact corpus of common words.

Let $P = \{p_1, p_2, \ldots, p_{48}\}$ be the complete set of English phonemes. Given a training corpus of common words W_{train}, any word $w_{\text{train}} \in W_{\text{train}}$ can be decomposed into a phoneme sequence

$$\text{Seq}(w_{\text{train}}) = [p_{i_1}, p_{i_2}, \ldots, p_{i_m}]. \tag{20}$$

Its corresponding IMU spectrogram $S(w_{\text{train}}) \in \mathbb{R}^{H \times W}$ is split into a sequence of pixel columns

$$\text{Col}(w_{\text{train}}) = [c_{i_1}, c_{i_2}, \ldots, c_{i_m}], \tag{21}$$

where each $c_p \in \mathbb{R}^{H \times 1}$ represents the column with phoneme p.

A mapping $M : p \mapsto c_p$ is learned by minimizing the following reconstruction loss over the training corpus:

$$\mathcal{L}_{\text{map}} = \sum_{w_{\text{train}}} \left\| S(w_{\text{train}}) - \text{Concat}\big(M(p_{i_1}), \ldots, M(p_{i_m})\big) \right\|_2^2. \tag{22}$$

For any unseen word $w_{\text{new}} \in V \setminus W_{\text{train}}$, the construction proceeds as follows. First, the word is decomposed into its phoneme sequence $\text{Seq}(w_{\text{new}}) = [p_{j_1}, p_{j_2}, \ldots, p_{j_k}]$ using a standard transcription tool. Next, the corresponding pixel columns $c_{j_\ell} = M(p_{j_\ell})$ are retrieved from the learned mapping. Finally, the spectrogram is reconstructed by concatenating these columns:

$$S(w_{\text{new}}) = \text{Concat}(c_{j_1}, \ldots, c_{j_k}), \tag{23}$$

with zero-padding applied to maintain a fixed temporal window.

The reconstructed spectrogram $S(w_{\text{new}})$ preserves the discriminative, phoneme-level structural features essential for recognition. In practice, the similarity between the reconstructed spectrogram and the ground-truth spectrogram $S_{\text{gt}}(w_{\text{new}})$ remains high, consistently satisfying

$$\text{SSIM}\big(S(w_{\text{new}}), S_{\text{gt}}(w_{\text{new}})\big) \geq 0.92. \tag{24}$$

Thus, the phoneme-based pixel reorganization strategy provides a theoretically sound and empirically effective mechanism for open-vocabulary word recognition.

C Vulnerabilities Assessment of Digit Password

Our attack achieves superior performance in digit recognition, thereby significantly undermining the security of digit passwords that are widely deployed in application scenarios such as personal identification number (PIN) verification and bank authentication.

We perform a quantitative risk analysis. We compare two essential security metrics (i.e., the search space and password entropy) of digit passwords when our attack is absent vs. when it is deployed.

1. Search Space Compression. For a k-digit password, the effective search space shrinks drastically from 10^k to 5^k because the recognition results of both human-voiced and earphone-produced speech are consistently confined to the top-5 candidate digits. This compression greatly facilitates enumeration-based attacks.

2. Password Entropy Reduction. Password entropy per digit is defined as:

$$H_i = -\sum_j p_j \log_2 p_j. \tag{25}$$

Without our attack, assuming uniform distribution over ten digits, each position has an entropy of approximately 3.3219 bits.

2.1 Human-voiced Digit Recognition: As shown in Figure 9(a), our attack achieves 75% top-1, 98% top-3, and 100% top-5 accuracy for human speech, constraining the effective space to the top-5 candidates. The resulting per-digit probability distribution is non-uniform: the top-1 digit accounts for 0.75, the second and third share 0.23 (0.115 each), and the fourth and fifth split the remaining 0.02 (0.01 each). Based on this distribution, the per-digit entropy is 1.1584 bits.

2.2 Earphone-produced Digit Recognition: For earphone-played digits, the attack attains 62.9% top-1, 91.7% top-3, and 99.1% top-5 accuracy. Using a similar non-uniform probability estimation constrained to the top-5 candidates, the resulting per-digit entropy is 1.579 bits.

Both entropy values are notably lower than that of a uniformly distributed 5-digit candidate set (2.3219 bits per position) and far below the 3.3219 bits of an unconstrained 10-digit password.

For an 8-digit PIN in Section 8.5, the total entropy drops from 26.58 bits (no attack) to $1.1584 \times 8 = 9.27$ bits for human-voiced passwords and to $1.579 \times 8 = 12.63$ bits for earphone-produced passwords. This reduction directly reflects a decline in security caused by the skewed recognition probabilities introduced by our attack.