Device-independent Smartphone Eavesdropping Jointly using Accelerometer and Gyroscope

Ming Gao, *Student Member, IEEE,* Yajie Liu, Yike Chen, Yimin Li, Zhongjie Ba, *Member, IEEE,* Xian Xu, Jinsong Han, *Senior Member, IEEE,* and Kui Ren, *Fellow, IEEE*

Abstract—Eavesdropping via inertial measurement units (IMUs) has brought growing concerns over smartphone users' privacy. In such attacks, adversaries utilize IMUs, including accelerometers and gyroscopes, which require zero permissions for access to acquire speeches. A common countermeasure is to limit sampling rates (within 200 Hz) to reduce overlap of vocal fundamental bands (85∼255 Hz) and inertial measurements (0∼100 Hz). Nevertheless, we observe that IMUs sampling below 200 Hz still record adequate speech-related information because of aliasing distortions. Accordingly, we propose a practical side-channel attack, namely *InertiEAR*, to break the defense of sampling rate restriction on the zero-permission eavesdropping. It leverages accelerometers and gyroscopes jointly to eavesdrop on both top and bottom speakers in smartphones. We exploit coherence between responses of the built-in accelerometer and gyroscope using a mathematical model. The coherence allows precise segmentation without manual assistance. We also mitigate the impact of hardware diversity and achieve better device-independent performance than existing approaches that have to massively increase training data from different smartphones for a scalable network model. These two advantages re-enable zero-permission attacks but also extend the attacking surface and endangering degree to off-the-shelf smartphones. *InertiEAR* achieves the recognition accuracy of 78.8% with the cross-device accuracy of up to 60.9% among 12 smartphones.

Index Terms—IMU eavesdropping, speech privacy, side channel, device-independence

1 Introduction

Privacy has always been a pivotal issue. People express increasing concerns over privacy protection, especially over eavesdropping via smartphones. Various sensors in smartphones intelligently gather information from the real world. However, those sensors risk malicious abuse. To resist privacy leakage, individuals consciously perform rigorous access control over explicitly privacy-related sensors such as microphones, cameras, and GPS.

Different from these sensitive sensors that are by default to the high permission level, built-in inertial measurement units (IMUs) are commonly regarded as the ones with low risk. Accessing IMUs requires little or zero permission. However, such sensors have been reported to facilitate so-called 'zero-permission' attacks to speech privacy [1], [2], [3], [4], [5], [6]. In such attacks, adversaries can access built-in accelerometers without users' permission nor attention. These IMUs can pick up speech signals from the on-board loudspeakers in the same smartphone. With a high sampling rate, IMUs are competent to cover the human voice's fundamental frequency band (85∼255 Hz) [7]. State-of-the-art (SOTA) attacks [2], [3] are able to obtain the alarming accuracy on speech recognition of 81% and speaker identification of 78%. Such threats have alerted the industry. A widely-held belief is to limit IMUs' sampling rates for avoiding that the range of inertial measurements overlaps with vocal fundamental bands. The risk of private speech leakage via zero-permission eavesdropping seems minimized. Under this common sense, Google has placed a restriction on IMUs where their sampling rate should not exceed 200 Hz [8].

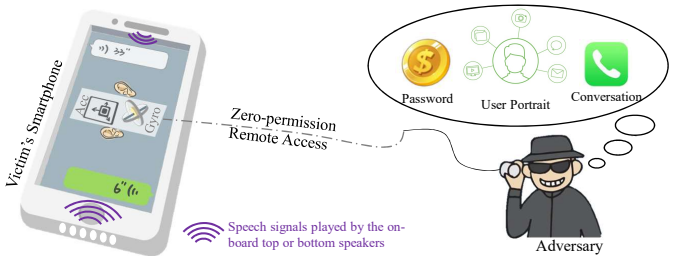


Fig. 1. *InertiEAR* allows the zero-permission attack for smartphone eavesdropping using IMUs with a limited sampling rate within 200 Hz.

Is this countermeasure effective against zero-permission eavesdropping? Experimentally, we observe IMUs still perform private speech theft even given the above restriction. Part of the high-frequency components in the user's voice would fall into low-frequency bands, namely aliasing distortions [9]. This indicates the possibility of recovering speech from residues contained in inertial readings sampled within 200 Hz. Taking a commercial off-the-shelf (COTS) smartphone, HUAWEI P40 as an example, its accelerometer can respond to audio signals of up to 6 kHz. It demonstrates smartphones are still vulnerable to zero-permission eavesdropping if merely restricting IMUs' sampling rates.

We further expand the attack surface to eavesdrop on the top and bottom on-board speakers. As illustrated in

- *Ming Gao, Yajie Liu, Yike Chen, Yimin Li, Zhongjie Ba, Jinsong Han, and Kui Ren are with the College of Computer Science and Technology, Zhejiang University and the ZJU-Hangzhou Global Scientific and Technological Innovation Center, China, 310000.*
  Email: gaomingppm@zju.edu.cn, yajie@zju.edu.cn, chenyike@zju.edu.cn, ninalym13@gmail.com, zhongjieba@zju.edu.cn, hanjinsong@zju.edu.cn, kuiren@zju.edu.cn.
  Xian Xu is with College of Civil Engineering and Architecture, Zhejiang University, China, 310000.
  Email: xianxu@zju.edu.cn.

Fig. 1, the IMU responds to audio signals emitted from both of the speakers. However, the top one is usually ignored by SOTA attacks [2], [3]. We jointly leverage accelerometers and gyroscopes in IMUs to aggravate privacy leakage from these speakers. Under such aggravation, adversaries can retrieve speech information emitted from any speaker in a smartphone, e.g., calls, audio media, and responses of voice assistants (VAs) that may mention locations and schedules.

To exploit the practice of eavesdropping, we further address two-fold realistic challenges that remain open in prior zero-permission attacks. (a) *Automation.* Previous approaches are lacking error-free signal segmentation methods. Traditional audio detection and segmentation techniques [10] hardly handle the additional noise in inertial data, especially under motion interference. Gyrophone [1] absolutely relies on manual divisions [1], while recent attacks count on filters to eliminate the influence of noise and human movement. But their effect is incomplete so that the segmentation is not precise (82% in [3] and 92% in [2]). In case of wrong divisions, manual inspection is inevitable. Apparently, such manual and error-prone segmentation cannot afford satisfactory speech recognition accuracy. (b) *Device-independence.* Recent zero-permission attacks [2], [3] improve the recognition accuracy by leveraging AI techniques. Nevertheless, they depend heavily on the training data and hence perform badly toward unseen smartphones because of the significant diversity of hardware features. Unfortunately, it is extremely difficult to construct a generalized network model based on training data collected from finite smartphone models. Meanwhile, adversaries cannot include all models of smartphones into training data due to the rapid increasing of smartphones. It is just in 2021 that 487 new smartphone models hit the market [11]. The adversaries have to know the smartphone models in advance and spend costly overhead in training thousand of specialized neural network for each model. Therefore, prior attacks are unscalable in terms of device-independent eavesdropping.

Accordingly, we develop a novel and practical attack, *InertiEAR*. It exploits the speaker-to-IMU side channel for eavesdropping on speeches from both top and bottom speakers in a smartphone. In particular, we address the limitation of previous work from the perspectives of automatic segmentation and device independence. We leverage the coherence between speech-related readings of the accelerometer and gyroscope. By the aid of a multiplier, we migrate these coherent responses into the direct-current bias, such that the responses are significantly distinguished from silent fragments in spite of noise and motion. Therefore, it supports an error-free segmentation without manual assistance. Meanwhile, we model hardware diversity of smartphones for enabling cross-device attacks. Our method integrates a range of techniques to eliminate the influence of hardware diversity and promote the device-independence from the perspective of data processing. We adopt DenseNet [12] for training a speech recognition model over the processed data and achieve high recognition accuracy of 78.8%. Using a trained model, *InertiEAR* supports an excellent performance with cross-device accuracy of 60.9%. Extensive evaluations on 12 COTS smartphones validate the effectiveness of *InertiEAR* under real-world scenarios. As a countermeasure, we propose defending methods against such eavesdropping without hardware modification.

In summary, our contributions are listed as follows:

- We revisit the threat of IMU-based eavesdropping and realize a side channel attack that breaks the restriction on sampling rates. A mathematical model is proposed to expand its attack surface and promote its practicality. We have released our collected data [13] [1] to facilitate the IMU-based eavesdropping research and appeal to public attention and countermeasures.
- We develop the automatic eavesdropping without manual assistance by the aid of accurate segmentation. By investigating inertial readings' coherence, our segmentation is error-free upon noise and motion interference.
- *InertiEAR* accomplishes a device-independent eavesdropping attack. Different from prior work, we suppress the hardware diversity by processing with a mathematical model rather than simply increasing training data, and hence significantly reduce the overhead of cross-device attacks.

2 IMUs and Their Sensitivity to Speech

An IMU is composed of a micro electromechanical system (MEMS) accelerometer and a MEMS gyroscope. The former measures acceleration and the latter supplies angular velocity. They directly contact the board where speakers lie in close proximity in a smartphone. Hence, speech signals emitted by speakers, both the top and bottom ones, inevitably leak into IMU's measurements.

Recent work has proved that IMUs are sensitive to speeches [1], [2], [3], [4], [5], [6]. Michalevsky et al. [1] study the effect of speeches on gyroscopes using independent loudspeakers placed on a common surface. They utilized multiple gyroscopes to capture speech vibration to obtain a high sampling rate. It reaches the quite low accuracy on recognition (26%) and speaker identification (50% among 10 speakers). Anand et al. [6] revisit IMUs' threat to private speeches under different scenarios, including human- and machine-rendered speeches travelling through the air or a common solid surface. They conclude that IMUs are only sensitive to signals propagating via solid with high power. Ba et al. [2] access built-in accelerometers to eavesdrop on the loudspeaker in a smartphone. With up to 500 Hz sampling rates, they achieve 70% accuracy on speaker identification and 78% accuracy on speech recognition. Anand et al. [3] slightly sharpen performances to 79% and 81% respectively but utilize accelerometers sampling at 4 kHz.

3 Threat Model

We assume that an adversary aims at private speeches emitted by speakers in the victim's smartphone. It threatens the security of remote calls and exposes other privacy (e.g., daily schedules, contacts, habits, and locations) via VAs' responses, personalized answers, and navigation services. Personal habits can be inferred from audio media for personalized advertising. Here, we define the adversary's capabilities as follows.

1. We have performed necessary data desensitization before the release according to the requirements of IBR and relevant regulations.

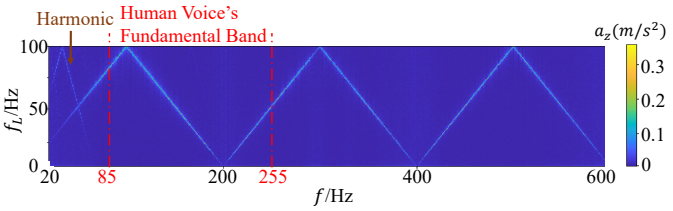


Fig. 2. Frequency responses of a smartphone's accelerometer. It can receive signals within the human voice's fundamental band.

Sensors Access. The adversary has installed a spy App on the victim's smartphone, under a mask of any legal App. It has no access to sensitive sensors like microphones but continuously captures IMU readings without permission. In an ideal attack, the spy App continuously captures inertial readings for stealing speeches. However, both Android and iOS limit the execution time of background applications with restricted access to inertial sensors [14], [15]. In this case, the adversary should declare a ForegroundService, resulting in a persistent notification that would make the spy App's activity noticeable. Though in presence of such limitations, the spy app can pose as a pedometer App. Therefore, it can obtain inertial readings in the background without increasing users' doubt even with a persistent notification.

Sampling Rate Limitation. The spy App runs at the highest available sampling rate. However, those of IMUs are limited to below 200 Hz by default [8] for privacy concern.

Attack Scenarios. The adversary can analyze speech-related IMU readings using several smartphones in advance for recognizing pre-trained sensitive words. It eavesdrops on the target smartphone's top and bottom speakers constantly. The target smartphone could be stationary or moving, typically on a table or in the user's hand. Fortunately, it might not be familiar with the target smartphone's model.

4 Model and Analysis

We explain the existence of the speaker-to-IMU side channel, followed by an analysis of noise and hardware diversity.

4.1 Speaker-to-IMU Side Channel

The accelerometer and gyroscope in a built-in IMU observe on-board speakers in a smartphone simultaneously, using three respective channels (i.e., axes) as follows,

$$\begin{aligned} \boldsymbol{A}(t) &= k_l \cdot M \cdot \boldsymbol{H_a} \cdot S_m(t) + \boldsymbol{N_a}, \\ \boldsymbol{G}(t) &= k_l \cdot M \cdot \boldsymbol{H_g} \cdot S_m(t) + \boldsymbol{N_g}, \end{aligned} \tag{1}$$

where $\boldsymbol{A}(t) = [a_x(t)\ a_y(t)\ a_z(t)]^T$ and $\boldsymbol{G}(t) = [\omega_x(t)\ \omega_y(t)\ \omega_z(t)]^T$ are IMU readings without noise and motion interference, $a_j(t)$ and $\omega_j(t)$ $(j = x, y, z)$ are readings of the accelerometer and gyroscope's corresponding axis, k_l is the level of volume setting decided by users, M is the highest volume of speakers, $\boldsymbol{H_i} = [h_{ix}\ h_{iy}\ h_{iz}]^T$ $(i = a, g)$ are 1×3 vectors with gain coefficients h_{ij}, $S_m(t)$ $(m = 1, 2)$ are speech signals emitted by the top and bottom speakers respectively, and $\boldsymbol{N_i}$ $(i = a, g)$ are channel noises. We mark the 2-norm $||\boldsymbol{H_i}||$ and the direction vector $\hat{\boldsymbol{H}}_{\boldsymbol{i}}$ as follows,

$$||\boldsymbol{H_i}||(t) = \sqrt{h_{ix}^2(t) + h_{iy}^2(t) + h_{iz}^2(t)},\ \hat{\boldsymbol{H}}_{\boldsymbol{i}} = \frac{\boldsymbol{H_i}}{||\boldsymbol{H_i}||}. \tag{2}$$

TABLE 1
SNR (dB) of IMU's response under each volume setting

Volume Setting		Bottom Speaker			Top Speaker		
		20%	60%	100%	20%	60%	100%
Acc	a_x	0.69	2.21	3.07	3.66	12.34	15.77
	a_y	4.24	5.49	5.88	11.17	19.90	23.32
	a_z	4.84	5.07	5.19	12.98	21.73	25.45
Gyro	ω_x	-7.66	-4.28	-6.18	-5.00	-0.31	2.69
	ω_y	-7.01	-5.04	-5.63	-6.71	-2.05	0.31
	ω_z	-6.70	-6.42	-5.56	-6.93	-5.39	-5.67

In an IMU whose sampling rate Fs is set below 200 Hz, an ideal low pass filter (LPF) should remove high-frequency components exceeding 100 Hz. In actual, because of the wide transition bandwidth of the LPF, these components are attenuated sightly rather than blocked entirely [16]. Components of the high frequency f are distorted into the low-frequency band f_L. Such a phenomenon, namely aliasing distortion, follows

$$f_L = ||f - n \times Fs||,\ (f_L < Fs/2,\ n \in \mathbb{N}). \tag{3}$$

The aliasing distortion and insecure filters are to blame for leaking private speech into IMUs.

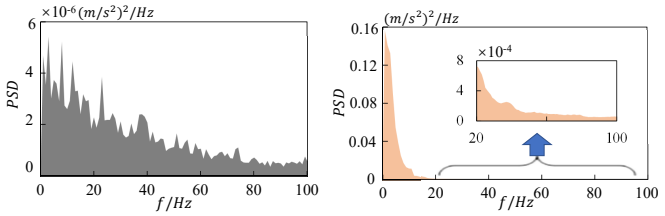
We conduct benchmark experiments to validate the derived model and demonstrate the feasibility of zero-permission attacks. We play an single-tone sound using the bottom loudspeaker of a HUAWEI P40, at its highest volume. The smartphone is placed on a table. The frequency sweeps from 20 Hz to 8 kHz at an interval of 1 Hz. We record its IMU's reading sampled at 200 Hz. The accelerometer's responses on the Z-axis are partly illustrated in Fig. 2. It can pick up the aliased tones up to 6 kHz. Similarly, the gyroscope can receive signals within 800 Hz. This phenomenon remains significant, whether the smartphone is placed on the table or held by hand.

We further measure the IMU's responses to the on-board speakers at different volume levels using the signal to noise ratio (SNR) defined as follows,

$$SNR = 10log_{10}\frac{P(T) - P(N)}{P(N)}, \tag{4}$$

where $P(T)$ and $P(N)$ are powers of sensors' outputs with and without speech. To be specific, we play a single tone signal at 150 Hz, a common frequency in the human voice [7]. It is emitted by the HUAWEI P40's top and bottom speakers at 20%, 60%, and 100% of its highest volume respectively. Tab. 1 lists SNRs of responses of each axis in the IMU. All axes in the accelerometer sense speech signals, with positive SNRs of up to 25 dB. They follow an approximately fixed SNR difference among axes, inferring the generally fixed distribution of inter-axial acoustic energy. This reflects the stability of $\hat{\boldsymbol{H}}_{\boldsymbol{i}}$, which comes from the relative position between the IMU and speakers.

Though gyroscopes initiate speech eavesdropping [1], they are discarded in recent attacks due to the low significance in comparison with accelerometers [2], [3], [6]. It is commonly asserted that a gyroscope performs barely sensitively to surface vibrations due to the duty of rotation measurement. In contrast, an actual gyroscope suffers from



(a) Intrinsic noise (b) Motion

Fig. 3. Noise energy distribution.

shock and vibration due to hardware defects [17]. Therefore, gyroscopes are able to pick up speech-related signals from surface vibrations, with an SNR up to 2.69 dB (See Tab. 1). Though with low SNRs in most settings, we further exploit and swell contained speech-related signals in Sec. 5.1.

In addition, these sensors show higher sensitivity to top speakers. Though they occupy the lower acoustic intensity, the closer proximity to the built-in IMU contributes to this phenomenon. Such high sensitivity leads to a new attack surface where zero-permission attacks can steal a wealth of private speeches from the top speakers. It lifts the unpractical restriction that victims have to turn up loudspeakers' volume to hear private speeches in SOTA attacks.

Due to the coincident observing location and asynchronous sampling, we characterize the accelerometer and gyroscope in an IMU as following two fundamental features: (a) *Coherence.* Their readings, originated from the same speech, share the identical frequency and phase. Such coherence can emphasize speech-related features for the error-free segmentation. (b) *Spectral expansion.* Considering their relative time-skew, we can combine them after normalization for a broader band [1].

4.2 Noise Analysis

A variety of noise would obscure speech-related signals in practical. We divide the noise into four categories and investigate their distributions and effects.

4.2.1 Intrinsic noise

We simplify intrinsic noise as a direct-current (DC) bias and an additive white noise [18]. The former can be removed by a high pass filter (HPF) directly, while the latter injects irregular power into each band. The white noise on each axis shares the identical distribution. On account of the white noise, simple high or low pass filters cannot suppress the effect of intrinsic noise, particularly on word segmentation.

4.2.2 Motion interference

Motion, especially human activities, exerts a dramatic effect on inertial measurements. These motion signals would overlap or even cover speech-related signals both in the accelerometer and the gyroscope. Fortunately, such interference concentrates on the low-frequency band. We recruit 16 volunteers aged from 18 to 50 for collecting motion data. They are required to install an APP that records their own smartphones' IMU readings[2] sampled at 200 Hz lasting two weeks. They are also instructed to avoid using on-board speakers during experiments. The collected data cover volunteers' daily motion, e.g., walking, running, bicycling, and driving. After removing stationary fragments, we present the statistical results of the collected human activities in Fig. 3(b). Although 98.20% of the energy is distributed below 20 Hz and 99.77% of that is within 80 Hz, there remains 0.23% of energy in the high-frequency band.

4.2.3 Harmonic

Ba et al. [2] point out the existence of surface vibration in an accelerometer. In some cases, external vibrations would bring about high-frequency pulses in accelerometer readings. We attribute such noise to harmonics. Recalling Fig. 2, tones swept from 20 to 60 Hz inject singles of the identical frequencies accompanied by additional third harmonics. We repeat this experiment where the smartphone is placed on a soft and sound-absorbing material, and the third harmonics disappear. Therefore, low-frequency vibrations of solid surfaces (e.g., tables) would distort accelerometer's readings with the harmonic energy leaked into the vocal fundamental band. Note that such harmonics exists only in accelerometers, but is absent in gyroscopes.

4.2.4 Ambient noise

Ambient noise falls into two categories, one around the target smartphone and the other around the remote caller. The former noise has been discussed thoroughly in the existing literature [2], [6], where it barely affects inertial readings. As for the latter one, it distorts speech signals from the acoustic point of view rather than the inertial one. Such noise varies in different environments. It burdens the adversarial segmentation and eavesdropping along with users' hearing. In this case, remote callers and victims are probable to increase volume actively. Furthermore, adversaries might collect noise distributions under different acoustic conditions and accordingly remove the noise influence using statistic-based noise cancellation methods [19].

In short, the aforementioned kinds of noise would affect inertia-based eavesdropping synthetically. The channel noise can be rewritten as follows

$$\boldsymbol{N_i} = \boldsymbol{B}(t) + \boldsymbol{N}_w(t) + \boldsymbol{M}(t) + \boldsymbol{N}_h(t),\ i = a, g \qquad (5)$$

where $\boldsymbol{B}(t)$ is the DC bias, $\boldsymbol{N}_w(t)$ is intrinsic white noise, $\boldsymbol{M}(t)$ is motion interference, and $\boldsymbol{N}_h(t)$ is the third harmonic noise but equals to 0 in a gyroscope.

To obtain clear speech-related data, $\boldsymbol{B}(t)$ and the low-frequency parts of $\boldsymbol{N}_w(t)$, $\boldsymbol{M}(t)$, and $\boldsymbol{N}_h(t)$ can be removed by an HPF. Although leaving slight influence on the adversarial speech recognition [2], the remnant components, such as short-time pulses, would nullify the effectiveness of statistic-based segmentation methods, e.g., absolute magnitude [2] and root mean square [3]. Instead, we propose an efficient solution in Sec. 5.2 based on the coherence of IMUs.

2. All experiments in this paper have obtained the IRB approval and we explicitly inform volunteers of the purpose behind the data. Here, these data are merely used for motion energy statistics, without any threaten to speech eavesdropping nor other privacy leakage. Devices include HUAWEI P20, P30, P40, Mate 10, Mate 20, Mi 8, Mi 10, HONOR 20, 30, OPPO Reno 5, Vivo S9, and Samsung Galaxy Note 20.

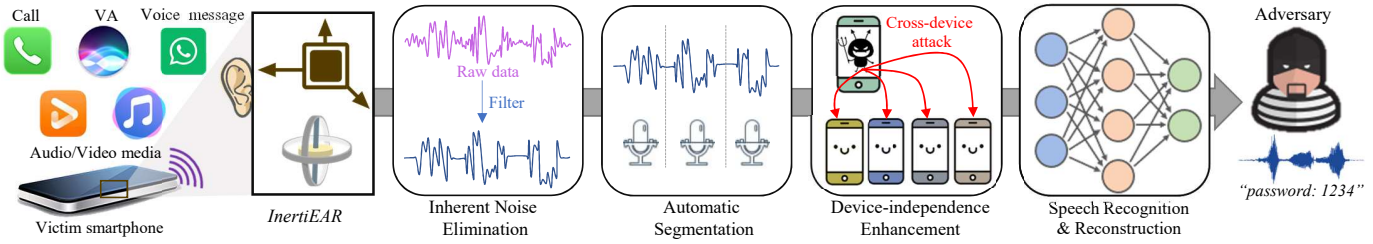


Fig. 4. *InertiEAR*, a practical zero-permission attack on smartphones, where an IMU is utilized for the on-board speakers eavesdropping even upon the recent limitation on sampling rates of below 200 Hz. It threats private speech, such as calls, audio media, voice messages, and VAs' responses.

4.3 Hardware Diversity

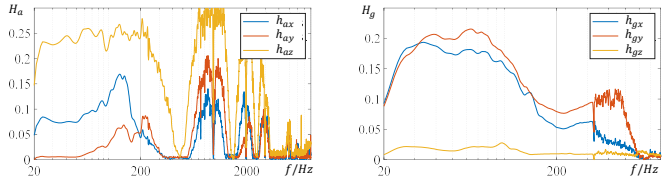
Diversity of hardware features is the key factor to impede the device-independent attack. These features will be remembered by trained network models for speech recovery, degrading their scalability. Here, we investigate sources of hardware diversity for further effect suppression.

Intrinsic noise $\boldsymbol{N}_w$: Speakers and IMUs own their unique hardware errors. Attendant intrinsic noise varies considerably among smartphones [20], [21].

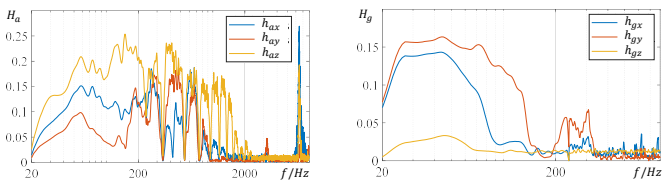
Response intensity M: Acoustic intensity determines the total energy of speech-related responses. Smartphones vary in the speaker power supply and perform differently even at the same volume level. Consequently, each built-in IMU has the distinctive response intensity.

Axial energy rate $\hat{\boldsymbol{H}}_{\boldsymbol{i}}$: Locations of the built-in IMU and speakers and their relative position are multifarious. Such diversity differentiates the proportion of speech-related energy among axes. For example, in a HUAWEI P40's accelerometer, the dominant axis, Z-axis, occupies about 50% of total energy, while Z-axis in a Samsung Galaxy S8 accounts for 59.4%, but in an Honour V30, X-axis dominates in some bands (about 100 to 300 Hz).

Frequency responses $\boldsymbol{H}(f)$: Hardware diversity would affect gain coefficients under inputs of different frequencies. On the one hand, hardware differences encourage diverse inherent frequency responses of speakers and IMUs. Responses of speaker-to-IMU side channels further combine the diversity of respective ones. On the other hand, inner LPFs introduce additional attenuation. Although failing in the complete removal, they still suppress out-of-band signals to a certain extent. Such effects depend on the parameter selection of LPFs with a perceived difference among IMUs. Moreover, a smartphone itself acts as an LPF [18] when speech signals propagate inside. Because of various sizes and masses, their filtering effects are diverse. The above causes contribute to complex and irregular responses, with those of two smartphones demonstrated in Fig. 5. Fortunately, sensors are designed for a stable response to in-band signals while LPFs are insensitive to low-frequency signals. The low-frequency responses are relatively smooth and flat (especially between 80 Hz and 200 Hz). In short, the primary distinction of frequency responses lies in the high-frequency distortion.



(a) HUAWEI P40



(b) Samsung Galaxy S8

Fig. 5. Frequency responses of two COTS smartphones.

Sampling rate Fs: Recalling Eq. 3, the sampling rate determines the aliasing distortion. There is a minor discrepancy in sampling rate among smartphones [2]. It indicates that the same out-of-band speech signals would fall into different bands in different IMUs. It further exacerbates differentiation in high-frequency bands among smartphones.

In conclusion, adversaries should remove intrinsic additive noise, eliminate axial energy difference, normalize response intensity, and mitigate high-frequency distortions. It is necessary to suppress the hardware diversity for device-independent attacks with better cross-device performance.

5 Attack Design

We propose a practical side-channel attack that utilizes the sensitivity of IMUs to speech signals emitted by on-board speakers for smartphone eavesdropping. It involves combined efforts from four modules, as illustrated in Fig. 4.

5.1 Intrinsic Noise Elimination

Intrinsic noise results in the low SNRs in Tab. 1 especially of gyroscopes at a low volume. Moreover, its diversity contributes to poor cross-device performance. We apply a wiener filter [19] to reduce such intrinsic noise, which aims at generalized stationary noise of a known distribution.

Adversaries can estimate the intrinsic noise distribution by collecting inertial readings when the smartphone is stationary without external inputs, for example, at midnight. Such a method demands no additional prior knowledge, e.g., the smartphone model. We conduct the wiener filtering on a HUAWEI P40 using the noise distribution. Resultant SNRs are increased by over 10 dB experimentally. In particular, even the SNR of the gyroscope's X-axis at the 20% volume (lowest one in Tab. 1) has increased to 7.11 dB after being filtered. It improves the significance of speech-related

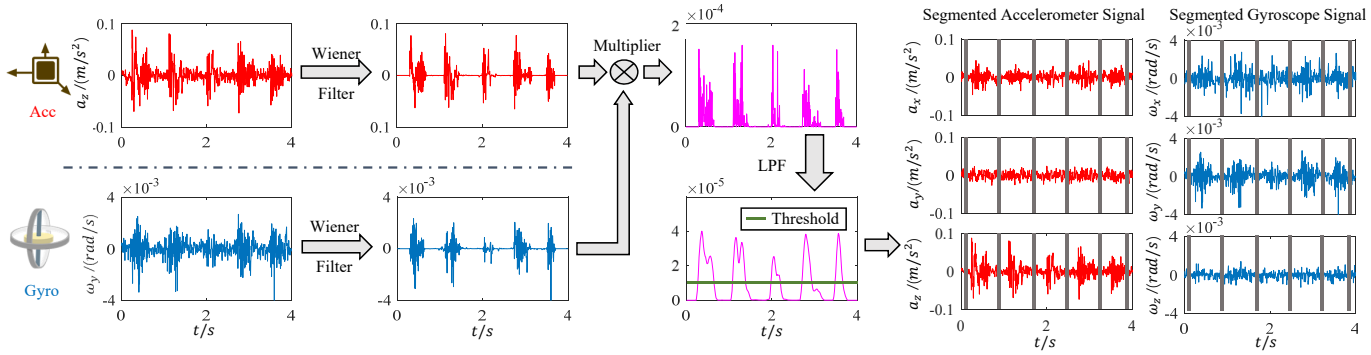


Fig. 6. An example of automatic segmentation that leverages IMU's coherence and distinguishes responses from silent fragments.

signals for the following segmentation and recognition. In addition, such a noise elimination technique also can be adopted to reduce the ambient noise around remote callers mentioned in Sec. 4.2.4 and experimental results are presented in Sec. 6.5.2.

5.2 Automatic Segmentation

An error-free and automatic segmentation technique is fundamental for practical eavesdropping. Otherwise, manual inspection and correction are inevitable but laborious. We exploit the coherence of the accelerometer and gyroscope and accordingly suppress noise and motion interference.

As mentioned in Sec. 4.1, the accelerometer and gyroscope in an IMU share the coherent readings. Specifically, they follow the identical frequency and a fixed phase difference. Conversely, the residual high-frequency components of the noise and motion are irrelevant interference between the accelerometer and gyroscope. Noise among sensors differs in spectrum distribution, while the acceleration and angular velocity describe motion from different perspectives and naturally are mutually independent. They barely overlap in the time and frequency domain simultaneously.

Under the above observation, we adopt a multiplier to stress speech-related signals. It migrates coherent components into the DC band with the second harmonics. These harmonics will be removed along with noise by an LPF. We suppose a single-frequency tone $sin(2\pi ft)$ to illustrate its effectiveness. In detail, we select inertial readings with the maximum energy among axes, e.g., $a_z(t)$ and $g_x(t)$ typically, and upsample them to 1000 Hz by linear interpolation to align time stamps. Such interpolation does not increase information nor relieve aliasing distortions. Followed by an LPF with the cut-off frequency of 20 Hz for removing intrinsic DC bias noise and low-frequency motion components, Eq. 5 is rewritten as follows,

$$\begin{aligned} a_z(t) &= k_a sin(2\pi f_L t) + n_{waz}(t) + m_{az}(t) + n_{haz}(t), \\ g_x(t) &= k_g sin(2\pi f_L t) + n_{wgx}(t) + m_{gx}(t), \end{aligned} \tag{6}$$

where k_i $(i = a, g)$ are the gain coefficients, and $n_{waz}(t)$, $n_{wgx}(t)$, $m_{az}(t)$, and $m_{gx}(t)$ are remnants of intrinsic white noise and motion on corresponding axes in the high-frequency band, and $n_{haz}(t)$ is the third harmonic noise in the accelerometer. After a multiplier, we have

$$a_z(t) \times g_x(t) = \frac{k_a k_g}{2} + \frac{k_a k_g sin(4\pi f_L t)}{2} + others, \tag{7}$$

where the latter two items will be removed by a LPF. The DC bias $\frac{k_a k_g}{2}$ significantly differentiates coherent response with non-vocal noise. Experimentally, we obtain an average DC bias of 1.77×10^{-5} in collected inertial data detailed in Sec 6.1. The biases maintain the same order of magnitude among various devices and settings. In comparison, the average result of intrinsic noise among 14 experimental smartphones keeps 1.3 $\times10^{-8}$ with a peak of 2.6×10^{-7} merely, and that of motion in Sec. 4.2.2 is 4.5×10^{-7} on average and at most 3.68×10^{-6}. In practice, we adjust Otsu algorithm [22] to decide thresholds for speech detection and segmentation in case that an extremely high outlier contributes to a high weighted threshold and the subsequent segment loss. We move each pair of threshold-crossing points by $\frac{Fs}{5}$ samples forward and backward respectively as cutting points. Fig. 6 illustrates a sample of signal segmentation. Note that all above processes in this subsection are used for calculating cutting points for segmentation, but not applied for following parts.

5.3 Device Independence Enhancement

For a practical eavesdropping attack with better cross-device performance, we remove device-dependent features caused by hardware diversity by processing. Following a wiener filter that has removed intrinsic noise in Sec. 5.1, we focus on axial energy rate, response intensity, and high-frequency distortions.

Dimension reduction. According to Eq. 1, axial energy differences $\hat{\boldsymbol{H}}_{\boldsymbol{i}}$ are redundant. They are directly related to the relative position between the IMU and speakers, rather than the one-dimensional speech signals. However, it may cost potential information loss to focus on only one axis but abandon others. Instead, we define

$$A^{\dagger}(t) = sign(a_{max}(t))||\boldsymbol{A}||(t), \tag{8}$$

where $sign(\cdot)$ is the sign function and $a_{max}(t)$ is the speech-related signal with the maximum energy among axes. We adopt $A^{\dagger}(t)$ rather than $||\boldsymbol{A}||$ to prevent frequency distortions. $G^{\dagger}(t)$ follows the same definition. This method maximizes multi-axial utilization and eliminates axial energy differences.

Normalization. We normalize $A^{\dagger}(t)$ and $G^{\dagger}(t)$ into [0,1]. This eliminates impacts of acoustic intensity, including speaker power M and volume settings k_l. It also converts readings of accelerometers and gyroscopes to a unified

dimension. Here, we concatenate them chronologically according to respective time stamps. Therefore, we double effective sampling rates and broaden the bandwidth of the speaker-to-IMU channel from 100 Hz to 200 Hz according to the Nyquist sampling theorem.

High-frequency suppression. High-frequency signals are folded into low bands. They are induced by aliasing and cannot be separated using digital approaches without hardware modification. In addition, out-of-band signals still contain information because of the vocal fundamental band (85∼255 Hz). In this case, we first exploit a HPF filter with a cut-off frequency of 80 Hz. It removes most of low-frequency motion within 80 Hz along with high-frequency noise of above 320 Hz that aliased into low bands. Rather than further separating high-frequency distortion, we delete samples randomly and downsample normalized signals into 390 Hz. It induces two-fold advantages. First, it eliminates the sampling rate differences among smartphones. Second, it aggravates high-frequency distortions [23] and obscures original features brought by hardware diversity, although reducing bandwidth to 195 Hz. Such random sampling deletions act as sampling jitters [23], leading to the following attenuation,

$$SNR = -20log_{10}(2\pi f \times rms(T_a)), \tag{9}$$

where $rms(T_a)$ is the aperture uncertainty caused by random downsampling. It sharply degrades high-frequency responses but induces few adverse effects on in-band signals.

5.4 Speech Recognition and Reconstruction

We design two networks to recognize and reconstruct the speech signals from the inertial readings.

Recognition. Processed inertial segmentation is transformed into 244×244 gray spectrogram images and fed to a DenseNet [12] for adversarial speech recognition. It establishes a dense connection between all the previous layers to the layers behind, and hence realizes feature reuse for less computational cost and better performance. We choose the cross-entropy as the training loss and use a piecewise momentum optimizer to optimize the model with a dropout rate of 0.3 during training.

Reconstruction. We generate 512×128 speech spectrograms from 128×128 gray inertial signal spectrograms using a network. The proposed reconstruction network consists of a 3-convolutional-layer encoder, five residual blocks, and a 3-deconvolutional-layer decoder. The encoder consists of three convolutional layers with 32 kernels of size $9 \times 9 \times 3$, 64 kernels of size $3 \times 3 \times 32$, and 128 kernels of size $3 \times 3 \times 64$. Correspondingly, the decoder consists of three deconvolutional layers with 64 kernels of size $3 \times 3 \times 128$, 32 kernels of size $3 \times 3 \times 64$, and 4 kernels of size $9 \times 9 \times 32$. We choose the L_1 training loss [24] and use a momentum optimizer with a time-based decay on the learning rate. Then we adopt Griffin-Lim algorithm [25] to estimate speech signals from the speech spectrograms. It is an iterative algorithm with two steps in each iteration: the first step is to modify the short time Fourier transform (STFT) of the estimated signal to be the same as the reconstructed speech spectrograms; the second step is to find an estimated signal whose STFT is close to the modefied STFT. After multiple steps of iterations, the final estimated signal is output as the reconstructed speech.

6 Evaluation

We conduct *InertiEAR* on COTS smartphones, and evaluate its performance through extensive real-world experiments.

6.1 Setup and Dataset

Audio dataset. We choose AudioMNIST dataset [26] that comprises 30 k single-digit audios from 60 speakers. The audio is played successively at an interval of 0.1 s. In addition, we recruit 6 volunteers (3 females and 3 males) to read 10 digits and 26 letters ten times at their average speech rate, around 110 words per minute (WPM) to serve as a homemade speech dataset.

IMU readings collection. We play speech signals using on-board top and bottom speakers respectively when target smartphones are placed on a table or held by hands. A spy App collects IMU readings sampled at 200 Hz by default in the background. The collected inertial data are randomly divided into two parts: 80% for training and 20% for testing. We mainly test on three smartphones: Samsung Galaxy S8, Google Pixel 4 (Android), and HUAWEI P40 (HarmonyOS). Additional 9 smartphones (including an iOS iPhone 11) are employed to test cross-device performance.

Metrics. We adopt top-k accuracy to assess the performance of speech recognition. It is defined as the ratio of the label within the top k classes correctly predicted. We introduce the mean square error to describe the difference between the reconstructed speech spectrogram $\hat{\boldsymbol{x}}$ and the ground truth speech spectrogram $\boldsymbol{x}$. It is calculated by $\frac{\sum_i^N (\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i)^2}{N}$, where i is the index of pixels and N is the total of pixels in each spectrogram.

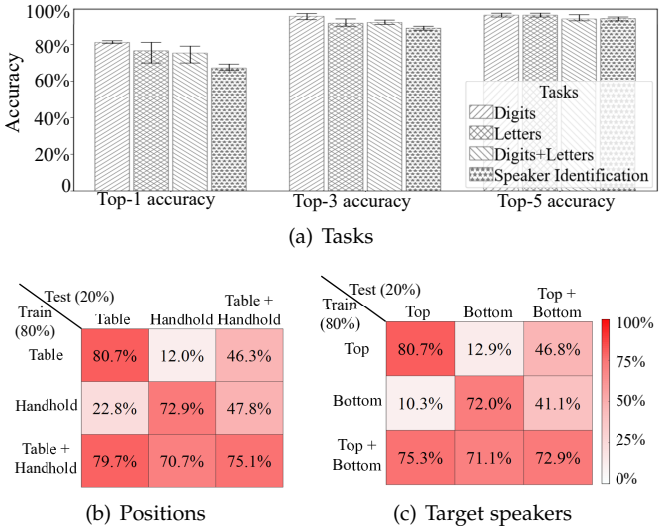
6.2 Overall Performance

InertiEAR brings great threats to speech privacy even given the limitation on sampling rate. It yields a 100% segmentation success rate and 78.8% recognition accuracy on average.

6.2.1 Segmentation

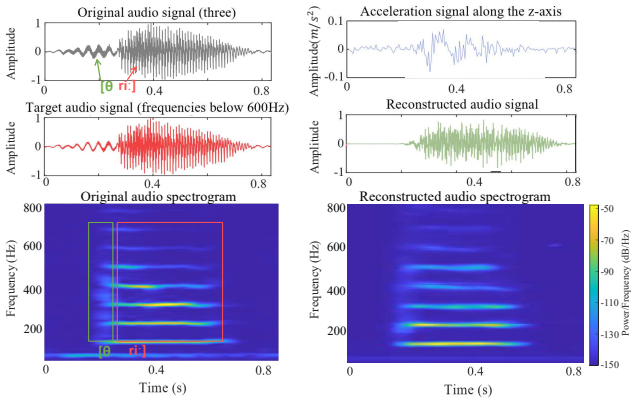
We develop the automatic segmentation with a success rate of up to 100% on audios composed of digits, letters, or a mixture. It works efficiently whether smartphones are placed on a table or held in the hand.

We take the influence of speech speed on segmentation into consideration. Volunteers repeat recording at three speeds: slow (below 95 WPM), average (around 110 WPM), and fast (over 130 WPM). *InertiEAR* succeeds to segment inertial data at the former two speeds. As for fast speed, it detects all fragments while the segmentation success rate shows a little drop of 1.38%. We find that the origin of error fragments lies in the liaisons where a volunteer speaks at a rate of above 160 WPM temporarily. Such a fast speed is not common in daily life or among VAs, and people usually slow down when sharing important information (e.g., password). Therefore, the proposed method entitles error-free segmentation in real-world scenarios. It supports a practical eavesdropping attack without manual assistant or correction that SOTA attacks require.



(a) Tasks (b) Positions (c) Target speakers

Fig. 7. Performance of speech recognition under different conditions.



(a) Audio signal (b) Reconstructed signal

Fig. 8. Comparison between original and reconstructed signals.

6.2.2 Recognition

We present *InertiEAR*'s performance under different conditions given the limitation on the sampling rate of 200 Hz. Fig. 7(a) shows rates of successful inferences from inertial readings. Remarkably, the digit recognition accuracy of *InertiEAR* even surpasses that of AcclEve [2] sampling at 500 Hz (78%) and approaches that of Spearphone [3] sampling at 4 kHz (81%). In addition, we implement it on an iPhone 11, and collect inertial data via a malicious web sampling at merely 60 Hz. *InertiEAR* maintains the top-1 digit-recognition accuracy of 43.7%. We take the initiative in realizing IMU-based eavesdropping on iOS-based smartphones, and verify the popularity of such zero-permission attacks among COTS smartphones.

We further study the feasibility of zero-permission attacks under different conditions. A Samsung Galaxy S8, for example, is placed on a table (labeled as 'Table') and held in users' hand (labeled as 'Handhold') respectively. Fig. 7(b) shows its testing digit-recognition accuracy when the smartphone is placed on a table (labeled as 'Table') and held in users' hand (labeled as 'Handhold'). The samples collected for 'Table' and 'Handhold' are 30,000 respectively. Though it performs badly (below 25%) when trained by data from merely one set but tested on data from the other, *InertiEAR* maintains the high recognition accuracy of over 70% when trained on both sets (labeled as 'Table+Handhold'). In particular, to simulate the most common posture of phone calls, we ask the 6 volunteers to hold smartphones not just in their hands, but also at their ears. The speeches are played by the top speakers. We collect 30,000 samples. *InertiEAR* still performs well in this scenario, with the top-1/3/5 accuracy of 71.2%, 92.4%, and 98.1% respectively. This result demonstrates the practical threat from *InertiEAR*. In short, different hand gestures and different smartphone orientations have little effect on eavesdropping.

Furthermore, we investigate the speech leakage of the top and bottom speakers. Fig. 7(c) demonstrates the threat of *InertiEAR* on them. Contrary to the common sense that top speakers should be more secure with the lower power, they risk the worse speech information leakage. The closer acoustic propagation distance through the smartphone is to blame for the vulnerability of top speakers. It exposes a new attack surface to zero-permission attacks.

6.2.3 Reconstruction

We reconstruct the audio signals using inertial data, as illustrated in Fig.8. Due to the limited sampling rate of inertial sensors (within 200 Hz), 800 Hz is almost the upper band of frequency that could be reconstructed. Compared with the target low-frequency component (the second row in Fig.8(a)) of a single-digit ('three') audio signal (the first row), the inertial signals (the first row in Fig.8(b)) has similar structures but is different in details. Nevertheless, we successfully reconstruct audio signals (the second row in Fig.8(b)) and achieve a low mean square error of around 4×10^{-4} on average. The comparison of spectrograms in the third row of Fig.8 also demonstrates the feasibility of the speech reconstruction. Although high-frequency consonant information may be lost to a certain extent, the reconstructed audio signals are still recognizable and intelligible. We recruit 16 volunteers aged 18 to 50. They are asked to listen to recognize reconstructed signals. All volunteers recognize the reconstructed signals correctly though they claim some fragments are a little unclear. The loss of true phases in the reconstructed signals is to blame. In future work, we will extract the phase information from inertial data for better speech reconstruction.

6.3 Scalability Study

We explicate the influence of device diversity in Sec. 4.3 and provide corresponding solutions. We verify the device independence of *InertiEAR* by testing the trained models using digits inertial data from other 10 unseen smartphones after processes in Sec. 5.3. As depicted in Fig. 9, we reach the superior cross-device performance of 33.1% on average, with a peak of 49.8%, using a model merely trained on data from two smartphones, almost twice than AccelEve [2] (of at most 26%). Even using the model trained on either a Samsung Galaxy S8 or a HUAWEI P40, *InertiEAR*'s cross-device performance still peaks at 44.1%. Our proposed approaches indeed prepare *InertiEAR* for the device-independent attack.

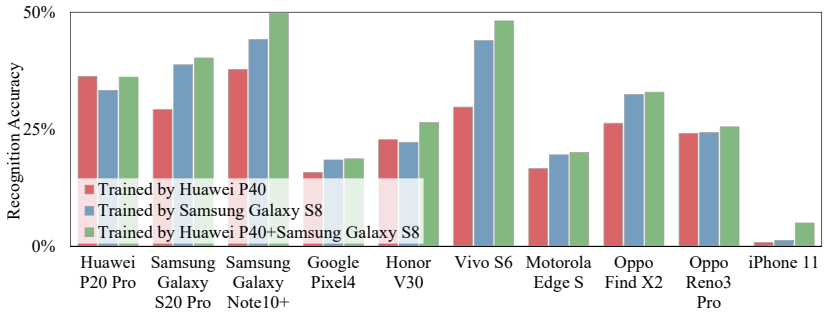


Fig. 9. Cross-device recognition accuracy using trained models.

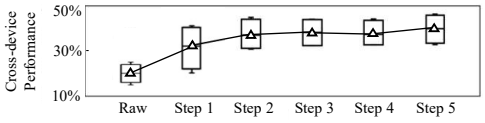


Fig. 10. Intermediate results show stepwise scalability improvement.

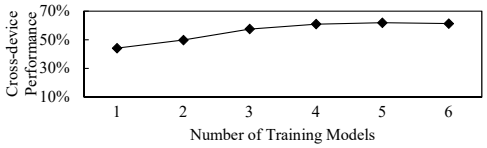


Fig. 11. Performance trained on multiple models.

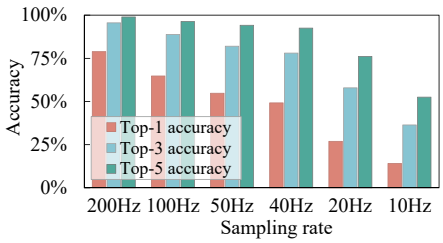


Fig. 12. Impact of sampling rate.

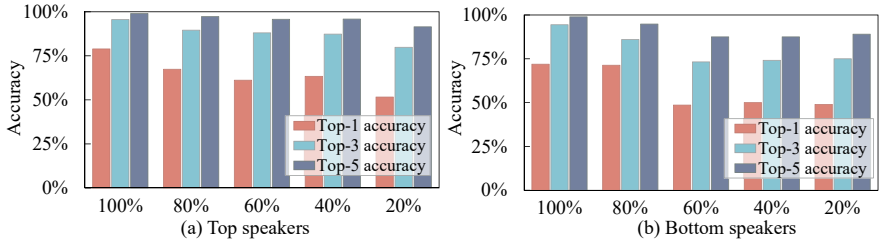


Fig. 13. Impact of volume setting.

We validate the effectiveness of each step mentioned in Sec. 5.3 for the device independence enhancement. We use intermediate data after each process from HUAWEI P40 to generate a recognition model respectively. We first train a model using raw triaxial accelerometer readings sampled at 200 Hz. Step 1 to 5 represent that data are processed by the wiener filtering, dimension reduction, normalization, concatenation with gyroscope readings, and downsampling successively. As shown in Fig. 10, the cross-device digit-recognition accuracy shows an upward trend as each means involves. Generally speaking, each process allows an improvement in cross-device performance.

We further exploit the potentials for improving the cross-device performance. In Fig. 9, the model trained by two smartphones is superior to those trained by one. This observation is correspond with the common sense is that the cross-device performance increases as the model is trained by the data from more smartphone models [2]. We explore the influence of the number of models involved in training, with results shown in Fig. 11. We succeed in an accuracy of up to 62.1% to eavesdropping on an unseen smartphone using 6 smartphone models. In particular, the cross-device performance reaches 60.9% when 4 models are involved and it improves extremely slightly as the number of training models increases. By comparison, without our proposed device-independent processes, a network trained on 10 smartphone models merely recognizes at most 37.6% of audios from an unseen smartphone experimentally. In practice, the low cost of attack preparation (i.e., training on just 4 models) is available for an average adversary and supports device-independent attacks with 60.9% accuracy.

6.4 Impact of Sampling Rate

The above experiments have confirmed the vulnerability of smartphones, even if sampling rates of built-in IMUs are imposed on the limitation of 200 Hz. To study the defending effectiveness of a restricted sampling rate against zero-permission eavesdropping attacks, we further burden *InertiEAR* using lower sampling rates. As illustrated in Fig. 12, the performance of our proposed attack deteriorates as the sampling rate falls. However, even the limitation of 40 Hz sampling rate is still at risk. *InertiEAR* maintains the top-1 accuracy of 49.2% and the top-5 accuracy of exceeding 90%. It breaks down the boundary on sampling rates that are expected to constrain IMU-based eavesdropping.

To explore the low bound of the sampling rate that may leak speech information, we further reduce the sampling rate to 10 Hz. The results show that the attack would be completely impractical with the top-5 accuracy of 52% (in comparison, that of randomly guessing is $1\text{-}0.9^5$=41%) when the sampling rates are reduced to 10 Hz. If the smartphones limit sampling rates within such a low value, the performance of motion-based applications (e.g., optical image stabilizers) will significantly degrade without the support from the high-accuracy motion measurement of inertial sensors. Therefore, the simple countermeasure of sampling rate limitation cannot effectively defend users' speech information. Instead, we propose two defenses to protect users' privacy in Sec. 7.2.

6.5 Impact of Practical Factors

6.5.1 Impact of Volume Setting

We evaluate the robustness of *InertiEAR* under different volume settings, which determines the SNR of speech-related inertial signals. As shown in Fig. 13, *InertiEAR* distinguishes fewer digits as the volume reduces. Nevertheless, even given the worst conditions of the lowest volumes, it succeeds in recognizing half of digits on average. Moreover, it keeps the top-5 accuracy of at least 89%. This dramatically

shortens the overall password search space for adversaries. In addition, *InertiEAR* maintains 100% segmentation success rate, except a slight drop of 1.3% when the volumes of bottom speakers are 20%. Though switching off bottom loudspeakers may work as a compromise, IMUs still keep eavesdropping on top speakers despite volume settings.

6.5.2 Impact of Ambient Noise

We divide the ambient noise into two categories, one around remote callers and the other around target smartphones. We leverage a wiener filter to estimate the ambient noise for suppressing its influence around remote callers, under the assumption that adversaries could collect noise distributions in different acoustic conditions. We recruit volunteers and call them in six realistic scenarios with different noise levels: laboratory, mall, street side, bus station, and metro station with or without trains running. SNRs of the noisy audio signals are 35.8 dB, 13.3 dB, 19.1 dB, 14.3 dB, 5.6 dB, and 17.2 dB respectively in the above environments. The corresponding accurate rates of recognition are 77.3%, 66.1%, 71.7%, 64.9%, 58.5%, and 70.4% . In extremely noisy scenarios (i.e., scenario 5, where SNR is only about 5 dB), speeches are almost totally covered by noise and thus difficultly to recognized to not only *InertiEAR* but also humans experimentally. Nevertheless, *InertiEAR* is able to resist ambient noise with accuracy above 65% in most environments. In addition, we repeat experiments when the target smartphones are located at the above noisy environments. Results verify IMU-based eavesdropping is undisturbed by environmental noise around target smartphones.

6.5.3 Impact of Shell

In daily life, users may equip their smartphones with protective shells. Shells may affect acoustic propagation through smartphones. Here we collect inertial data from smartphones with original shells to test *InertiEAR* using models trained on smartphones without shells. These testing shells are provided together with smartphones by manufacturers. They are made of TPU rubber material and measure approximately 2 mm in thickness. *InertiEAR* keeps the average recognition accuracy of 70.5%. In spite of a slight drop, such accuracy still significantly threaten users' privacy.

6.6 End-to-end Attack Case Study

We conduct an end-to-end attack in password inference. Suppose that a victim requests a password from a remote caller, but the on-board speakers in the victim's smartphone is spied by *InertiEAR*. The adversary aims at locating and recognizing the password from IMU readings.

6.6.1 Word Recognition and Reconstruction

InertiEAR can also be adopted for recognizing and reconstructing words and sentences. We demonstrate its performance via a common scenario where users mention some hot words before private information, e.g., passwords.

We select 10 hot words (as listed in Fig. 14) to train a recognition model that is aimed at detecting and identifying these words from sentences. We choose the Speech commands dataset [27] and a homemade dataset to train the word recognition model. The homemade dataset includes

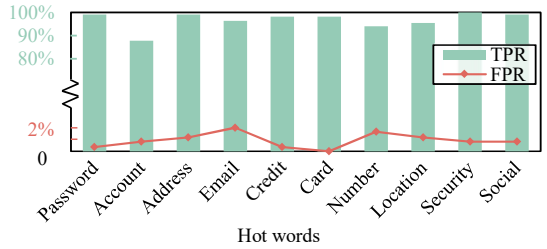


Fig. 14. TPR and FPR for word recognition.

TABLE 2
Accuracy of Password inference.

Setting	Segmentation Success Rate	Digit Recognition Accuracy		
		Top-1	Top-3	Top-5
On-table	97.8%	68.2%	90.1%	97.9%
Sitting	92.3%	66.2%	88.0%	97.5%
Walking	91.2%	61.7%	80.2%	90.6%

128×10 hot words and 2560 insensitive words (negative samples) from four volunteers (two males and two females). Here, the number of negative samples is 20 times that of each hot word. Therefore, we re-weight the losses to balance classes with a hyper-parameter α for the negative samples and 20α for the hot word samples. α is set 0.1 during training and we test on test sentences collected from the volunteers. As illustrated in Fig.14, *InertiEAR* is still effective on recognizing words and sentences. It achieves a true positive rate (TPR) of 96.7% on average and a false positive rate (FPR) below 2%. The more distinctive spectrograms of these hot words account for the better recognition accuracy in comparison with the digit recognition.

We further reconstruct the test sentences from the inertial data. The mean square error maintains below 7.5×10^{-4}. We also play reconstructed signals to volunteers and ask them to double check hot words labelled by the above recognition model. The TPR remains unchanged and the FPR declines to less than 0.5% after manual double check.

6.6.2 Password Inference

After detecting hot words, the adversary steal the following password via inertial data. We assume three scenarios where the target smartphone is placed on a table (labeled as 'On-table'), or held in the hand of a sitting (labeled as 'Sitting') or walking (labeled as 'Walking') victim. We recruit four volunteers (2 females and 2 males) acting as victims to hold a HUAWEI P40 with their preferred hand gestures and smartphone orientations. We recruit other four volunteers (2 females and 2 males) acting as the remote callers, each of which is asked to tell victims 20 random 8-digits passwords preceded by the hot word 'password (is)' via phone calls per scenario, followed by several non-digital voices (80 passwords in each scenario and 240 in total).

We first segment inertial readings and maintain a success rate of above 91%. A trained binary classifier is also leveraged to detect digits. Such a digit detection is more practical than a hot word search, considering that victims would not always prompt adversaries via specific words. As listed in Tab. 2, *InertiEAR* recognizes 60% of digits in passwords. It

affords a significant key space reduction in practical attacks on password eavesdropping. Results also demonstrate the robustness of *InertiEAR* against movement interference.

6.7 Comparison with SOTA Attacks

We compare the proposed attack, *InertiEAR*, with SOTA techniques [1], [2], [3] in Tab. 3. Gyrophone [1] initially studies speech recognition from gyroscopes of merely 26% accuracy. AccelEve [2] extends attacks onto smartphones' loudspeakers using 500 Hz sampling rate and promote the accuracy in speech recognition substantially. Spearphone [3] improves recognition and identification accuracy slightly, but demands 4 kHz sampling rate that is impractical especially after Google's updating [8]. Though with the lowest sampling rate, *InertiEAR* achieves the satisfactory performance with 78.8% recognition accuracy, and reconstruct speech signals accurately. Besides, SOTA attacks suffer from diversity of smartphone hardware for a generalized model. *InertiEAR* has no such issues instead, with the cross-device recognition accuracy of up to 60.9%, not to mention that it also has other advantages, such as error-free segmentation, high accuracy at low volume settings and robust performance under the motion interference.

7 Discussion

7.1 Further Improvement

We probe hardware diversity using a mathematical model and enable the device-independent eavesdropping with 60.9% cross-device recognition accuracy, but there is much room to be desired. We regard diverse frequency responses as an obstacle to the further improvement of zero-permission attacks. Firstly, we mitigate aliasing distortion using the random downsampling which, however, yields finite benefits. Although it fades out-of-band signals' characteristics, a learning-based model is still likely to exact and remember these features from aliasing components. Secondly, there are minor fluctuations in low-frequency responses. These fluctuations may also contribute to the device dependence. A potential solution for the adversary is to measure responses in the band of 85$\sim$200 Hz using smartphones of the same model in advance. This requires the knowledge about victims' smartphone models but costs less time to sweep single-frequency tones than collecting huge amounts of speech-related inertial data and training another new model. In addition, the material and thickness of shells equipped by target smartphones might exert an impact on the recognition accuracy, and we will discuss their influence in further work.

7.2 Countermeasure

We summarize existing defenses and propose practical methods with neither additional hardware modification nor inconvenience for users. We have reported the eavesdropping threat and potential countermeasures to related manufacturers. In particular, HUAWEI Inc. has adopted our suggestions and deploy our proposed solutions on their smartphones against such side-channel attacks.

7.2.1 Existing methods

Sampling rate limitation and secure filters: As illustrated in Sec. 6.4, the limitation on sensors' refreshing rate shows poor performance for speech privacy protection. The aliasing distortion and insecure filters are to blame. It is a plausible solution to using a secure analogy filter and implementing access control on IMUs. However, the former requires hardware modification on the filter circuit, while a low sampling rate and additional access control [2] on IMUs block their convenience and efficient perception.

Damping and isolating: Another idea is to shield built-in IMUs from speech signals. These sensors are expected to be isolated physically [3] or encircled by acoustic dampening materials [28]. However, these methods are unpractical particularly in mobile devices for additional modification, limited space, and cost.

7.2.2 Our solutions

Resonant noise: Although Android, iOS, and HarmonyOS do not provide users with on-off switches of inertial sensors, users are suggested to induce resonant noise proactively using on-board speakers to jam IMUs during speeches. These resonant acoustics, even at a low volume, can bring about significant noise into multiple axes simultaneously [9], [29], [30], [31]. Accelerometers in Samsung Galaxy S8, for instance, resonate with frequencies centered approximately 6.5 kHz in Fig. 5(b). This method blocks coherence-based segmentation and confuses recognition with miniature hearing interference on humans and no additional modification.

Oversampling: From the perspective of manufacturers, we propose a novel countermeasure at the system level. The failure of the sampling rate limitation and filters lies in the inevitable transition zone after the cut-off frequency in an analogy filter, where out-of-band signals would not be removed completely. Instead, inspired by the oversampling used in audio processing, we sample IMUs at their high rates to avoid the aliasing distortion, followed by a digital LPF to eliminate the high-frequency band over 85 Hz to remove human voices. Under this guidance, we test on a HUAWEI P40 that samples at 500 Hz and deploy a digital LPF (85 Hz cut-off frequency). Only 19.8% of audios can be recognized in this case. Therefore, the risk of privacy leakage from IMUs is significantly weakened.

8 Related Work

Privacy Leakage. Adversaries can access IMU in both iOS and Android without users' permission [1] for gather personal privacy. Apart from speech eavesdropping [1], [2], [3], [4], [5], [6], adversaries can maliciously leverage inertial data to infer victims' keystroke [32], [33], [34], [35], [36], [37] that may leak user password and PIN, and localize and trace victims [38], [39], [40], [41], [42]. Moreover, adversaries can identify devices using IMUs' unique hardware fingerprints for the personalized advertising [21], [43], [44], [45].

IMU-assisted Application. IMUs are widely deployed in various systems on users' convenience due to their sensitivity and low cost. Besides accurate attitude calculation and movement estimation [46], they can also support gesture recognition [47], [48], [49], [50], [51], sign language translation [52], covert channel communication [31], [53], [54] and

TABLE 3
Comparison with SOTA Attacks

Attack	Sensor	Sampling rate	Segmentation	Speech Recognition	Speech Reconstruction	Motion Robustness	Device Independence
Gyrophone [1]	Gyro.	200 Hz	Manually	26%	×	×	Not learning-based
AccelEve [2]	Acc.	500 Hz	92%	78%	✓	× Segmentation ✓ Recognition	at most 26% (trained on 2 models)
Spearphone [3]	Acc.	4 kHz	81%	78%	×	An HPF above 20 Hz but no evaluating	×
InertiEAR	Acc.+ Gyro.	within 200 Hz	100%	78.8%	✓	✓ Segmentation ✓ Recognition	60.9% (4 models) 49.8% (2 models)

behavior and biometric characteristics based authentication [55], [56], [57], [58], [59].

Spoofing Attacks on IMUs. The security and integrity of inertial data also have raised people's concerns. It has been reported that IMUs are vulnerable against acoustic interference [29]. Attackers leverage modulated acoustics to modify inertial data and therefore conduct denial of service (Dos) attacks [29], [60] or even manipulate IMU-based systems [9], [30], [61].

9 Conclusion

We realize *InertiEAR*, a practical speaker-to-IMU side channel attack. It breaks the restriction on sampling rates for smartphones eavesdropping. Both the automatic segmentation and device-independence promote the scalability of such zero-permission eavesdropping in reality, and appeal to people for necessary countermeasures to resist its threat.

Acknowledges

This paper is partially supported by the National Key R&D Program of China (2021QY0703), National Natural Science Foundation of China under grant U21A20462, 61872285, 62032021, 61772236, 62172359, and 61972348, Research Institute of Cyberspace Governance in Zhejiang University, Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (Grant No. 2018R01005), Zhejiang Key R&D Plan (Grant No. 2019C03133), Ant Group Funding No.Z51202000234, and Alibaba-Zhejiang University Joint Institute of Frontier Technologies.

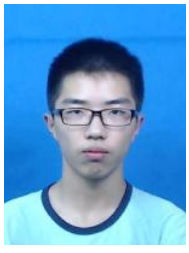
References

[1] Y. Michalevsky, D. Boneh, and G. Nakibly, "Gyrophone: Recognizing speech from gyroscope signals," in *USENIX Security Symposium*, 2014.
[2] Z. Ba, T. Zheng, X. Zhang, Z. Qin, B. Li, X. Liu, and K. Ren, "Learning-based practical smartphone eavesdropping with built-in accelerometer," in *NDSS*, 2020.
[3] S. A. Anand, C. Wang, J. Liu, N. Saxena, and Y. Chen, "Spearphone: A lightweight speech privacy exploit via accelerometer-sensed reverberations from smartphone loudspeakers," in *ACM WiSec*, 2021.
[4] L. Zhang, P. H. Pathak, M. Wu, Y. Zhao, and P. Mohapatra, "Accelword: Energy efficient hotword detection through accelerometer," in *ACM MobiSys*, 2015.
[5] J. Han, A. J. Chung, and P. Tague, "Pitchin: Eavesdropping via intelligible speech reconstruction using non-acoustic sensor fusion," in *ACM/IEEE IPSN*, 2017.
[6] S. A. Anand and N. Saxena, "Speechless: Analyzing the threat to speech privacy from smartphone motion sensors," in *IEEE S&P*, 2018.
[7] I. R. Titze and D. W. Martin, "Principles of voice production," *Journal of the Acoustical Society of America*, vol. 104, no. 3, pp. 1148–1148, 1998.
[8] Android for Developers, "Behavior changes: Apps targeting android 12," https://developer.android.com/about/versions/12/behavior-changes-12#motion-sensor-rate-limiting, 2021.
[9] Y. Tu, Z. Lin, I. Lee, and X. Hei, "Injected and delivered: Fabricating implicit control over actuation systems by spoofing inertial sensors," in *USENIX Security Symposium*, 2018.
[10] A. Yoshida, H. Mizuno, and K. Mano, "Segment selection method based on tonal validity evaluation using machine learning for concatenative speech synthesis," in *IEEE ICASSP*, 2008.
[11] Canalys Inc., "Smartphone analysis," https://www.canalys.com/analysis/smartphone, 2022.
[12] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE CVPR*, 2017.
[13] P4VEMENT, "Inertiear: Speech-related inertial date," https://github.com/P4VEMENT/InertiEAR$_{S}peech-related_{I}nertial_{D}ate.git$, 2022.
[14] Android for Developers, "Background," https://developer.android.com/about/versions/oreo/background, 2021.
[15] Apple Developers, "Preparing your ui to run in the background," https://developer.apple.com/documentation/uikit/app-and-environment/scenes/preparing-your-ui-to-run-in-the-background, 2021.
[16] TomRoelandts, "The transition bandwidth of a filter depends on the window type," https://tomroelandts.com/articles/the-/ transition-bandwidth-of-a-filter-depends-on-the-window-type, 2021.
[17] Analog Devices, Inc., "Shock and vibration rejection of mems gyroscopes," https://developer.android.com/about/versions/12/behavior-changes-12#motion-sensor-rate-limiting, 2021.
[18] Analog Devices,Inc., "Anticipating and managing critical noise sources in mems gyroscopes," https://www.analog.com/en/technical-articles/critical-noise-sources-mems-gyroscopes.html, 1999.
[19] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Wiley, 1949.
[20] Z. Zhou, W. Diao, X. Liu, and K. Zhang, "Acoustic fingerprinting revisited: Generate stable device ID stealthily with inaudible sound," in *ACM CCS*, 2014.
[21] J. Zhang, A. R. Beresford, and I. Sheret, "Sensorid: Sensor calibration fingerprinting for smartphones," in *IEEE S&P*, 2019.
[22] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
[23] B. Brannon and A. Barlow, "Aperture uncertainty and adc system performance," https://www.analog.com/media/en/technical-documentation/application-notes/an-501.pdf, 2006.
[24] R. Chen and I. C. Paschalidis, "A robust learning approach for regression models based on distributionally robust optimization," *The Journal of Machine Learning Research*, vol. 19, pp. 517–564, 2018.

[25] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.
[26] S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller, and W. Samek, "Interpreting and explaining deep neural networks for classification of audio signals," *CoRR*, vol. abs/1807.03418, 2018.
[27] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *CoRR*, vol. abs/1804.03209, 2018.
[28] R. Dean, N. Burch, M. Black, A. Beal, and G. Flowers, "Microfibrous metallic cloth for acoustic isolation of a mems gyroscope," *SPIE*, 2011.
[29] Y. Son, H. Shin, D. Kim, Y. Park, J. Noh, K. Choi, J. Choi, and Y. Kim, "Rocking drones with intentional sound noise on gyroscopic sensors," in *USENIX Security Symposium*, 2015.
[30] T. Trippel, O. Weisse, W. Xu, P. Honeyman, and K. Fu, "WALNUT: waging doubt on the integrity of MEMS accelerometers with acoustic injection attacks," in *IEEE EuroS&P*, 2017.
[31] M. Gao, F. Lin, W. Xu, M. Nuermaimaiti, J. Han, W. Xu, and K. Ren, "Deaf-aid: Mobile iot communication exploiting stealthy speaker-to-gyroscope channel," in *ACM MobiCom*, 2020.
[32] E. Miluzzo, A. Varshavsky, S. Balakrishnan, and R. R. Choudhury, "Tapprints: your finger taps have fingerprints," in *ACM MobiSys*, 2012.
[33] Z. Xu, K. Bai, and S. Zhu, "Taplogger: inferring user inputs on smartphone touchscreens using on-board motion sensors," in *ACM WiSec*, 2012.
[34] X. Liu, Z. Zhou, W. Diao, Z. Li, and K. Zhang, "When good becomes evil: Keystroke inference with smartwatch," in *ACM CCS*, 2015.
[35] C. Wang, X. Guo, Y. Wang, Y. Chen, and B. Liu, "Friend or foe?: Your wearable devices reveal your personal PIN," in *ACM ASIACCS*, 2016.
[36] E. Owusu, J. Han, S. Das, A. Perrig, and J. Zhang, "Accessory: password inference using accelerometers on smartphones," in *ACM HotMobile*, 2012.
[37] L. Cai and H. Chen, "Touchlogger: Inferring keystrokes on touch screen from smartphone motion," in *USENIX HotSec*, 2011.
[38] R. Gao, B. Zhou, F. Ye, and Y. Wang, "Knitter: Fast, resilient single-user indoor floor plan construction," in *IEEE INFOCOM*, 2017.
[39] A. Rai, K. K. Chintalapudi, V. N. Padmanabhan, and R. Sen, "Zee: Zero-effort crowdsourcing for indoor localization," in *ACM MobiCom*, 2012.
[40] F. Li, C. Zhao, G. Ding, J. Gong, C. Liu, and F. Zhao, "A reliable and accurate indoor localization method using phone inertial sensors," in *ACM UbiComp*, 2012.
[41] J. R. Kwapisz, G. M. Weiss, and S. Moore, "Activity recognition using cell phone accelerometers," *SIGKDD Explor.*, vol. 12, no. 2, pp. 74–82, 2010.
[42] D. Jung, T. Teixeira, and A. Savvides, "Towards cooperative localization of wearable sensors using accelerometers and cameras," in *IEEE INFOCOM*, 2010.
[43] H. Liu, X.-Y. Li, L. Zhang, Y. Xie, Z. Wu, Q. Dai, G. Chen, and C. Wan, "Finding the stars in the fireworks: Deep understanding of motion sensor fingerprint," in *IEEE INFOCOM*, 2018.
[44] S. Dey, N. Roy, W. Xu, R. R. Choudhury, and S. Nelakuditi, "Accelprint: Imperfections of accelerometers make smartphones trackable," in *NDSS*, 2014.
[45] Y. Son, J. Noh, J. Choi, and Y. Kim, "Gyrosfinger: Fingerprinting drones for location tracking based on the outputs of MEMS gyroscopes," *ACM TOPS*, vol. 21, no. 2, pp. 1–25, 2018.
[46] R. Quinonez, J. Giraldo, L. E. Salazar, E. Bauman, A. A. Cárdenas, and Z. Lin, "SAVIOR: securing autonomous vehicles with robust physical invariants," in *USENIX Security Symposium*, 2020.
[47] H. Wen, J. R. Rojas, and A. K. Dey, "Serendipity: Finger gesture recognition using an off-the-shelf smartwatch," in *ACM CHI*, 2016.
[48] G. Laput, R. Xiao, and C. Harrison, "Viband: High-fidelity bio-acoustic sensing using commodity smartwatch accelerometers," in *ACM UIST*, 2016.
[49] Y. Katsuhara and H. Kaji, "Towards multi-person motion forecasting: Imu based motion capture approach," in *UbiComp/ISWC*, 2019.
[50] P. Yang, L. Xie, C. Wang, and S. Lu, "Imu-kinect: A motion sensor-based gait monitoring system for intelligent healthcare," in *UbiComp/ISWC*, 2019.
[51] H. Aly and M. Youssef, "Zephyr: Ubiquitous accurate multi-sensor fusion-based respiratory rate estimation using smartphones," in *IEEE INFOCOM*, 2016.
[52] J. Hou, X.-Y. Li, P. Zhu, Z. Wang, Y. Wang, J. Qian, and P. Yang, "Signspeaker: A real-time, high-precision smartwatch-based sign language translator," in *ACM MobiCom*, 2019.
[53] N. Roy, M. Gowda, and R. R. Choudhury, "Ripple: Communicating through physical vibration," in *USENIX NSDI*, 2015.
[54] K. Block, S. Narain, and G. Noubir, "An autonomic and permissionless android covert channel," in *ACM WiSec*, 2017.
[55] C. Wu, K. He, J. Chen, Z. Zhao, and R. Du, "Liveness is not enough: Enhancing fingerprint authentication with behavioral biometrics to defeat puppet attacks," in *USENIX Security Symposium*, 2020.
[56] X. Xu, J. Yu, Y. chen, Q. Hua, Y. Zhu, Y.-C. Chen, and M. Li, "Touchpass: Towards behavior-irrelevant on-touch user authentication on smartphones leveraging vibrations," in *ACM MobiCom*, 2020.
[57] W. Chen, L. Chen, Y. Huang, X. Zhang, L. Wang, R. Ruby, and K. Wu, "Taprint: Secure text input for commodity smart wristbands," in *ACM MobiCom*, 2019.
[58] H. Feng, K. Fawaz, and K. G. Shin, "Continuous authentication for voice assistants," in *ACM MobiCom*, 2017.
[59] J. Liu, C. Wang, Y. Chen, and N. Saxena, "Vibwrite: Towards finger-input authentication on ubiquitous surfaces via physical vibration," in *ACM CSS*, 2017.
[60] Z. Wang, K. Wang, B. Yang, S. Li, and A. Pan, "Sonic gun to smart devices: Your devices lose control under ultrasound/sound," in *Blackhat USA*, 2017.
[61] X. Ji, Y. Cheng, Y. Zhang, K. Wang, C. Yan, W. Xu, and K. Fu, "Poltergeist: Acoustic adversarial machine learning against cameras and computer vision," in *IEEE SP*, 2021.

Ming Gao is a Ph.D. candidate at the school of cyber science and technology, Zhejiang University. He received the Master and Bachelor degree from Xi'an Jiaotong University. His research interests include cyber-physical security, mobile computing, and privacy protection. He is a recipient of the Best Paper Award Nomination from SenSys'21.

Yajie Liu is currently a postgraduate student with the School of Cyber Science and Technology, Zhejiang University.



Yike Chen is working toward the PhD degree at the School of Cyber Science and Technology, Zhejiang University. His research interests include mobile computing and smart sensing.

Yimin Li received her M.Sc. from University College London in 2021. She was a Visiting Student with Zhejiang University from 2019 to 2020. Her research interests include wireless sensor network and privacy protection.

Zhongjie Ba received the Ph.D. degree in computer science and engineering from the State University of New York at Buffalo in 2019. He is currently a ZJU100 Young Professor with the College of Computer Science and Technology and the Institute of Cyberspace Research (ICSR), Zhejiang University, Hangzhou, China. He was a Post-Doctoral Researcher with the School of Computer Science, McGill University. His current research interests include the security and privacy aspects of Internet of Things, artificial intelligence powered mobile sensing, and forensic analysis of multimedia contents.

Xian Xu is a professor at the College of Civil Engineering and Architecture, Zhejiang University. His research interests include smart structural health monitoring.

Jinsong Han received his Ph.D. degree in computer science from Hong Kong University of Science and Technology in 2007. He is now a professor at the School of Cyber Science and Technology, Zhejiang University. He is a senior member of the ACM and IEEE. His research interests focus on IoT security, smart sensing, wireless and mobile computing.

Kui Ren received the Ph.D. degree from the Worcester Polytechnic Institute, Worcester, MA, USA. He is currently a Professor of computer science and technology and the Director of the Institute of Cyberspace Research, Zhejiang University, Hangzhou, Zhejiang, China. His current research interests include cloud and outsourcing security, wireless and wearable system security, and artificial intelligence security. Dr. Ren is also a Distinguished Scientist and Fellow of the ACM. He was a recipient of the IEEE CISTC Technical Recognition Award 2017 and the NSF CAREER Award in 2011.