InertiEAR: Automatic and Device-independent IMU-based Eavesdropping on Smartphones

Ming Gao, Yajie Liu, Yike Chen, Yimin Li, Zhongjie Ba, Xian Xu, Jinsong Han*
Zhejiang University, China
{gaomingppm, yajie, chenyike, zhongjieba, xianxu, hanjinsong}@zju.edu.cn, ninalym13@gmail.com

***Abstract*—IMU-based eavesdropping has brought growing concerns over smartphone users' privacy. In such attacks, adversaries utilize IMUs that 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 experimentally 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, *InertiEAR*, to break the defense of sampling rate restriction on the zero-permission eavesdropping. It leverages IMUs to eavesdrop on both top and bottom speakers in smartphones. In the *InertiEAR* design, we exploit coherence between responses of the built-in accelerometer and gyroscope and their hardware diversity 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 a recognition accuracy of 78.8% with a cross-device accuracy of up to 49.8% among 12 smartphones.**

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

I. Introduction

Privacy has always been a pivotal issue during the information age. People express increasing concern 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]–[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].

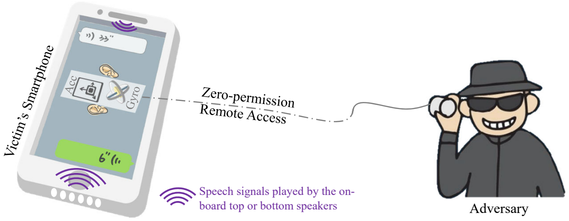


Fig. 1. *InertiEAR* allows the zero-permission attack for smartphone eavesdropping using IMUs with a limited sampling rate within 200 Hz.

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 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 daily 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

[*] Jinsong Han is the corresponding author.

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. For a certain smartphone unseen, the adversaries have to know it in advance and spend costly overhead in training a specialized neural network. 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 [11] for training a speech recognition model over the processed data and achieve a high recognition accuracy of 78.8%. Using a trained model, *InertiEAR* supports an excellent performance with cross-device accuracy of 49.8%. 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 *InertiEAR* that breaks the restriction on sampling rates. A mathematical model is proposed to expand its attack surface and promote its practicality.
- We develop the automatic eavesdropping without manual assistance by the aid of accurate segmentation. By thoroughly 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.

II. Background

A. IMUs and Their Sensitivity to Speech

An IMU embedded in a smartphone is composed of a 3-axis micro electromechanical system (MEMS) accelerometer and a 3-axis 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]–[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.

B. Related work

IMUs are widely deployed in various systems on users' convenience due to their sensitivity and low cost. Besides accurate attitude calculation and movement estimation [12], they can also support gesture recognition [13]–[17], sign language translation [18], covert channel communication [19]–[21] and behavior and biometric characteristics based authentication [22]–[26]. However, adversaries can access IMU in both iOS and Android without permission [1] to gather personal privacy, including speech [1]–[6], keystroke [27]–[32], localization [33]–[37], and device fingerprints [38]–[41].

III. 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.

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 victims' permission.

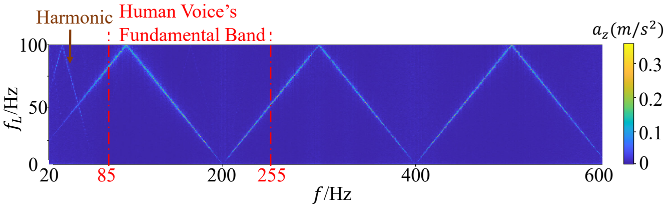


Fig. 2. Responses of a smartphone's accelerometer to frequency-sweeping tones. It can receive signals within the human voice's fundamental band.

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

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.

IV. Model and Analysis

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

A. 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,

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

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 [42]. Components of the high frequency f are distorted into the low-frequency band f_L. Such a phenomenon, namely aliasing, 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.

TABLE I
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

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 signal powers of sensors' outputs with and without the presence of 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. I lists SNRs of speech-related 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 shock and vibration due to hardware defects [43]. Therefore, gyroscopes are able to pick up speech-related signals from surface vibrations as well, with an SNR up to 2.69 dB (See Tab. I). Though with low SNRs in most settings, we further exploit and swell contained speech-related signals in Sec. V-A.

In addition, these sensors show higher sensitivity to the top speaker. 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].

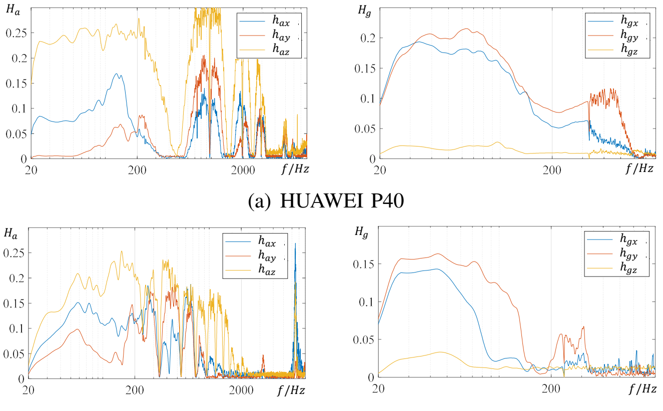
B. Noise Analysis

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

1) Intrinsic noise: We simplify intrinsic noise as a direct-current (DC) bias and an additive white noise [44]. 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.

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[1] 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. 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.

3) Harmonic: Ba et al. [2] point out the existence of surface vibration in an accelerometer. We attribute such noise to harmonics. Recalling Fig. 2, tones swept from 20 to 60 Hz inject sigles 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.



(a) HUAWEI P40

(b) Samsung Galaxy S8

Fig. 3. Frequency responses of two COTS smartphones.

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. We discuss one possible solution in Sec. VII.

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

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

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. V-B based on the coherence of IMUs.

C. 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 [41], [45].

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

[1] 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.

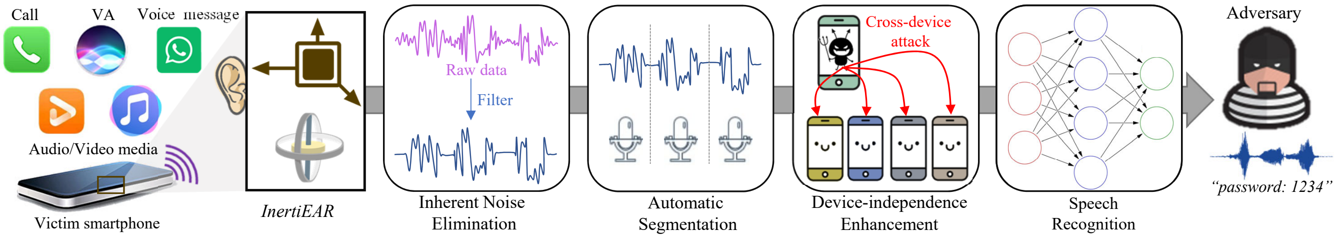


Fig. 4. *InertiEAR*, a practical zero-permission attack on smartphones, where an IMU is utilized for the on-board speakers eavesdropping even upon the recent limitation on sampling rates of below 200 Hz. It threats private speech, such as calls, audio media, voice messages, and VAs' responses.

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.

Frequency responses $H(f)$: 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 [44] 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. 3. 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 plat (especially between 80 Hz and 200 Hz). In short, the primary distinction of frequency responses lies in the high-frequency distortion.

Sampling rate Fs: Recalling Eq. 3, the sampling rate determines the aliasing distortion. In practice, 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 would further exacerbate 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.

V. 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.

A. Intrinsic Noise Elimination

Intrinsic noise results in the low SNRs in Tab. I especially of gyroscopes at a low volume. Moreover, its diversity contributes to poor cross-device performance. We apply a wiener filter [46] 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. I) has increased to 7.11 dB after being filtered. It improves the significance of speech-related signals for the following segmentation and recognition.

B. 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. IV-A, 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,

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

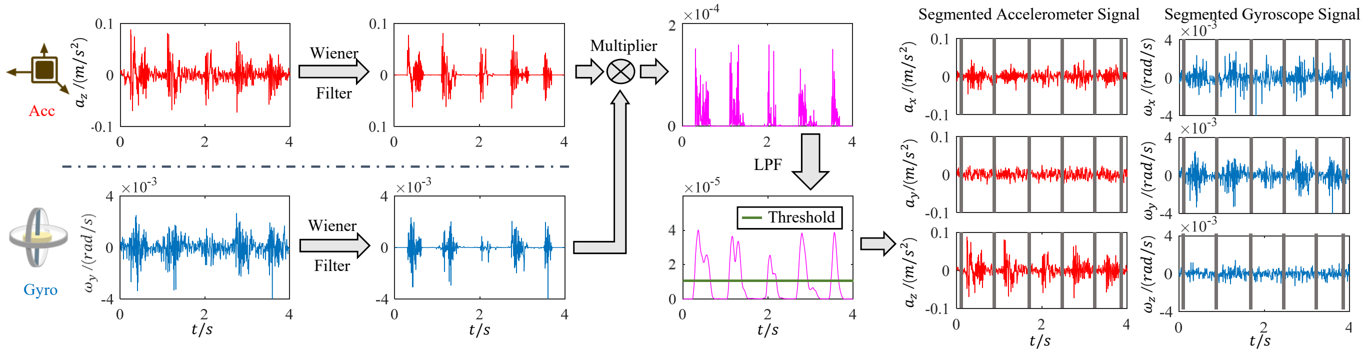


Fig. 5. An example of automatic segmentation that leverages IMU's coherence and distinguishes responses from silent fragments despite noise and motion.

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

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

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 VI-A. 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. IV-B2 is 4.5×10^{-7} on average and at most 3.68×10^{-6}. In practice, we adjust Otsu algorithm [47] 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. 5 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.

C. 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. V-A, 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

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

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.

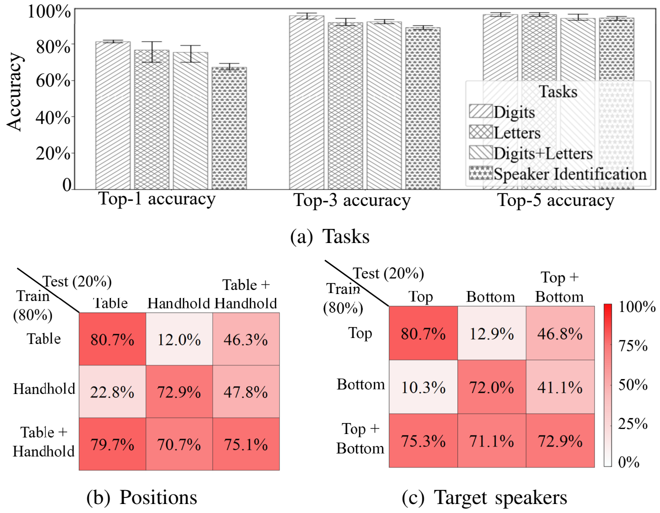
High-frequency suppression. High-frequency signals are folded into low bands. They are induced by aliasing and cannot be separated using digital approaches without hardware modification. In addition, out-of-band signals still contain information because of the vocal fundamental band (85-255 Hz). In this 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 [48] and obscures original features brought by hardware diversity, although reducing bandwidth to 195 Hz. Such random sampling deletions act sampling jitters [48], leading to the following attenuation,

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

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.

D. Speech Recognition

Processed inertial segmentation is transformed into 244 × 244 gray spectrogram-images and fed to a DenseNet [11] 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.



(a) Tasks

(b) Positions

(c) Target speakers

Fig. 6. Performance of speech recognition under different conditions.

VI. Evaluation

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

A. Setup and Dataset

Audio dataset. We choose AudioMNIST dataset [49] that comprises 10k single-digit audios from 20 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).

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-based), and HUAWEI P40 (HarmonyOS-based). Additional 9 smartphones (including an iOS-based iPhone 11) are employed to test cross-device performance.

B. 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.

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.

2) Recognition: We present *InertiEAR*'s performance under different conditions given the limitation on the sampling rate of 200 Hz. Fig. 6(a) shows rates of successful inferences from inertial readings within top-k predicted results. 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.

To 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. 6(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'). 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'). Furthermore, we investigate the speech leakage of the top and bottom speakers. Fig. 6(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.

C. Scalability Study

We explicate the influence of device diversity in Sec. IV-C 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. V-C. As depicted in Fig. 7, we reach the superior cross-device performance of 33.1% on average, with

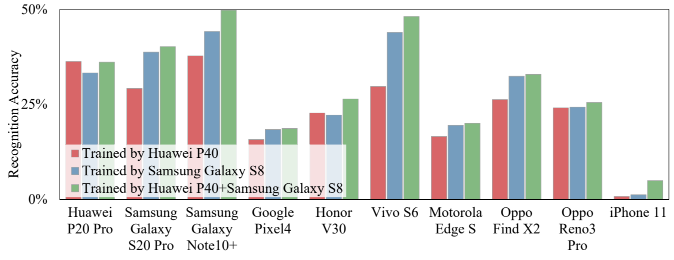


Fig. 7. Cross-device recognition accuracy using trained models.

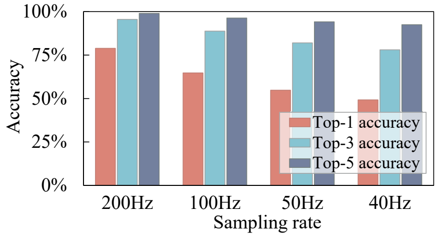


Fig. 8. Impact of sampling rate.

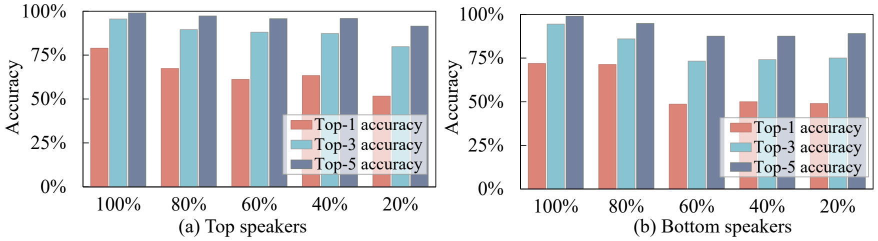


Fig. 9. Impact of volume setting.

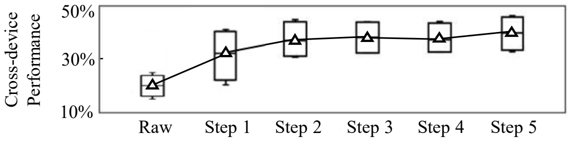


Fig. 10. Intermediate results show stepwise scalability improvement.

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.

We validate the effectiveness of each step mentioned in Sec. V-C for the device independence enhancement. We use intermediate data after each process from HUAWEI P40 to generate a recognition model respectively. We test it using data from other smartphones. As shown in Fig. 10[2], 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. The above results substantiate the device independence of our proposed attack.

D. 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. 8, 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 smartphones eavesdropping.

E. Impact of Volume Setting

We further evaluate the robustness of *InertiEAR* under different volume settings, which determines the SNR of speech-related inertial signals. As shown in Fig. 9, *InertiEAR* can distinguish fewer digits as the volume reduces. Nevertheless, even given the worst conditions of the lowest volumes, it succeeds in recognizing half of target digits on average. Moreover, it keeps the high 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 volume 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.

F. End-to-end Case Study: Password Inference

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. 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 the remote callers. Each of them are asked to tell us 20 random 8-digits passwords via phone calls using a HUAWEI P40 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 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. II, *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.

TABLE II
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%

G. Comparison with SOTA Attacks

We compare the proposed attack, *InertiEAR*, with SOTA techniques [1]–[3] in Tab. III. Gyrophone [1] initially studies speech recognition from gyroscopes of merely 26% accuracy.

[2] 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.

TABLE III
COMPARISON WITH SOTA ATTACKS

Attack	**Sensor**	**Sampling rate**	**Segmentation**	**Speech Recognition**	**Speaker Identification**	**Motion Robustness**	**Device Independence**
Gyrophone [1]	Gyroscope	200 Hz	Manually	26%	50%	×	Not learning-based
AccelEve [2]	Accelerometer	500 Hz	92% success rate	78%	70%	× Segmentation ✓ Recognition	at most 26%
Spearphone [3]	Accelerometer	4 kHz	82% success rate	81%	78%	An HPL above 20 Hz but no evaluating	×
InertiEAR	Accelerometer+ Gyroscope	within 200 Hz	100% success rate	78.8%	67.3%	✓ Segmentation ✓ Recognition	up to 49.8% 33.1% on average

AcclEve [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. 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 49.8%, not to mention that it also has other advantages, such as error-free segmentation, high accuracy at low volume settings and robustness against motion.

VII. DISCUSSION

A. Further Improvement

We probe hardware diversity using a mathematical model and enable the device-independent eavesdropping with 49.8% 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 85-200 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. As for ambient noise around remote callers in Sec. IV-B, advanced speech enhancement technologies [50] relieve such noise interference. Additionally, sophisticated adversaries may analyze the ambient noise distribution and eliminate it using wiener filters (See Sec. V-A).

B. 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.

1) Existing methods: **Sampling rate limitation and secure filters**: As illustrated in Sec. VI-D, 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. They are expected to be isolated physically [3] or encircled by acoustic dampening materials [51]. However, these methods are unpractical particularly in mobile devices for additional modification, space, and cost.

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], [20], [52], [53]. Accelerometers in Samsung Galaxy S8, for instance, resonate with frequencies centered approximately 6.5 kHz in Fig. 3(b). This method blocks coherence-based segmentation and confuses recognition with miniature hearing interference on humans and no additional modification.

VIII. 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.

ACKNOWLEDGMENT

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] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE CVPR*, 2017.

[12] 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.

[13] H. Wen, J. R. Rojas, and A. K. Dey, "Serendipity: Finger gesture recognition using an off-the-shelf smartwatch," in *ACM CHI*, 2016.

[14] G. Laput, R. Xiao, and C. Harrison, "Viband: High-fidelity bio-acoustic sensing using commodity smartwatch accelerometers," in *ACM UIST*, 2016.

[15] Y. Katsuhara and H. Kaji, "Towards multi-person motion forecasting: Imu based motion capture approach," in *UbiComp/ISWC*, 2019.

[16] 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.

[17] H. Aly and M. Youssef, "Zephyr: Ubiquitous accurate multi-sensor fusion-based respiratory rate estimation using smartphones," in *IEEE INFOCOM*, 2016.

[18] 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.

[19] N. Roy, M. Gowda, and R. R. Choudhury, "Ripple: Communicating through physical vibration," in *USENIX NSDI*, 2015.

[20] 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.

[21] K. Block, S. Narain, and G. Noubir, "An autonomic and permissionless android covert channel," in *ACM WiSec*, 2017.

[22] 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.

[23] 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.

[24] 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.

[25] H. Feng, K. Fawaz, and K. G. Shin, "Continuous authentication for voice assistants," in *ACM MobiCom*, 2017.

[26] J. Liu, C. Wang, Y. Chen, and N. Saxena, "Vibwrite: Towards finger-input authentication on ubiquitous surfaces via physical vibration," in *ACM CSS*, 2017.

[27] E. Miluzzo, A. Varshavsky, S. Balakrishnan, and R. R. Choudhury, "Tapprints: your finger taps have fingerprints," in *ACM MobiSys*, 2012.

[28] Z. Xu, K. Bai, and S. Zhu, "Taplogger: inferring user inputs on smartphone touchscreens using on-board motion sensors," in *ACM WiSec*, 2012.

[29] X. Liu, Z. Zhou, W. Diao, Z. Li, and K. Zhang, "When good becomes evil: Keystroke inference with smartwatch," in *ACM CCS*, 2015.

[30] 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.

[31] E. Owusu, J. Han, S. Das, A. Perrig, and J. Zhang, "Accessory: password inference using accelerometers on smartphones," in *ACM HotMobile*, 2012.

[32] L. Cai and H. Chen, "Touchlogger: Inferring keystrokes on touch screen from smartphone motion," in *USENIX HotSec*, 2011.

[33] R. Gao, B. Zhou, F. Ye, and Y. Wang, "Knitter: Fast, resilient single-user indoor floor plan construction," in *IEEE INFOCOM*, 2017.

[34] A. Rai, K. K. Chintalapudi, V. N. Padmanabhan, and R. Sen, "Zee: Zero-effort crowdsourcing for indoor localization," in *ACM MobiCom*, 2012.

[35] 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.

[36] 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.

[37] D. Jung, T. Teixeira, and A. Savvides, "Towards cooperative localization of wearable sensors using accelerometers and cameras," in *IEEE INFOCOM*, 2010.

[38] 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.

[39] S. Dey, N. Roy, W. Xu, R. R. Choudhury, and S. Nelakuditi, "Accelprint: Imperfections of accelerometers make smartphones trackable," in *NDSS*, 2014.

[40] 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.

[41] J. Zhang, A. R. Beresford, and I. Sheret, "Sensorid: Sensor calibration fingerprinting for smartphones," in *IEEE S&P*, 2019.

[42] 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.

[43] 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.

[44] 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.

[45] Z. Zhou, W. Diao, X. Liu, and K. Zhang, "Acoustic fingerprinting revisited: Generate stable device ID stealthily with inaudible sound," in *ACM CCS*, 2014.

[46] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Wiley, 1949.

[47] 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.

[48] 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.

[49] 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.

[50] K. Sun and X. Zhang, "Ultrase: single-channel speech enhancement using ultrasound," in *ACM MobiCom*, 2021.

[51] R. Dean, N. Burch, M. Black, A. Beal, and G. Flowers, "Microfibrous metallic cloth for acoustic isolation of a mems gyroscope," *SPIE*, 2011.

[52] 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.

[53] 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.