Spearphone: A Lightweight Speech Privacy Exploit via Accelerometer-Sensed Reverberations from Smartphone Loudspeakers

S Abhishek Anand
University of Alabama at Birmingham
anandab.official@live.com

Chen Wang
Louisiana State University
chenwang1@lsu.edu

Jian Liu
University of Tennessee, Knoxville
jliu@utk.edu

Nitesh Saxena
University of Alabama at Birmingham
saxena@uab.edu

Yingying Chen
Rutgers University
yingche@scarletmail.rutgers.edu

ABSTRACT

In this paper, we build a speech privacy attack that exploits *speech reverberations* from a *smartphone's inbuilt loudspeaker* captured via a zero-permission motion sensor (accelerometer). We design our attack *Spearphone*, and demonstrate that speech reverberations from inbuilt loudspeakers, at an appropriate loudness, can impact the accelerometer, leaking sensitive information about the speech. In particular, we show that by exploiting the affected accelerometer readings and carefully selecting feature sets along with off-the-shelf machine learning techniques, Spearphone can perform *gender classification* (accuracy over 90%) and *speaker identification* (accuracy over 80%) for the audio/video playback on the smartphone for our recorded dataset. We use lightweight classifiers and an off-the-shelf machine learning tool so that the attacking effort is minimized, making our attack practical. Our results with testing the attack on a voice call and voice assistant response were also encouraging, showcasing the impact of the proposed attack. In addition, we perform *speech recognition* and *speech reconstruction* to extract more information about the eavesdropped speech to an extent. Our work brings to light a fundamental design vulnerability in many currently-deployed smartphones, which may put people's speech privacy at risk while using the smartphone in the loudspeaker mode during *phone calls*, *media playback* or *voice assistant interactions*.

CCS CONCEPTS

• **Security and privacy → Side-channel analysis and countermeasures**;

KEYWORDS

side channel, speech privacy, motion sensors

ACM Reference Format:
S Abhishek Anand, Chen Wang, Jian Liu, Nitesh Saxena, and Yingying Chen. 2021. Spearphone: A Lightweight Speech Privacy Exploit via Accelerometer-Sensed Reverberations from Smartphone Loudspeakers. In *14th ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec '21), June 28-July 2, 2021, Abu Dhabi, United Arab Emirates.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3448300.3468499

1 INTRODUCTION

Today's smartphones contain a plethora of sensors aiming to provide a comprehensive and rich user experience. A known security vulnerability associated with smartphone motion sensors is their unrestricted access on most current mobile platforms (e.g., the Android OS), essentially making them *zero-permission* sensors. Recent research [10, 11, 16, 26–28] exploits motion sensors for eavesdropping on keystrokes, touch input and speech. Since the Android mobile operating system has a market share of 75.16% worldwide and 42.75% in the United States [4], this security vulnerability is of extreme concern especially in terms of speech privacy.

Expanding on this research line in significant ways, we investigate a new attack vulnerability in motion sensors that arises from the *co-located* speech source on the smartphone (smartphone's in-built loudspeakers). Our work exploits the motion sensors (accelerometer) of a smartphone to capture the speech reverberations (surface-aided and aerial) generated from the smartphone's loudspeaker while listening onto a voice call or any media in the loudspeaker mode. The reverberations are generated due to the smartphone's body vibrating due to forced vibrations [18], similar to a sounding board of a piano. Using this attack, we show that it is possible to compromise the speech privacy of a live human voice, without the need of recording and replaying it at a later time instant. As the threat of exploiting smartphone's loudspeaker privacy using motion sensor arises due to co-location of the speech source, it showcases the perils to a user's privacy in seemingly inconspicuous threat instances as described below:

- *Remote Caller's Speech Privacy Leakage in Voice Calls*: The proposed attack can eavesdrop on voice calls to compromise the speech privacy of a remote end user in the call. A smartphone's loudspeaker can leak the speech characteristics of a remote end party in a voice call via its motion sensors. These speech characteristics may be their gender, identity or the spoken words during the call (by performing speech recognition or reconstruction).

WiSec '21, June 28-July 2, 2021, Abu Dhabi, United Arab Emirates

ACM ISBN 978-1-4503-8349-3/21/06…$15.00
https://doi.org/10.1145/3448300.3468499

- *Speech Media Privacy Leakage:* In the proposed attack, on-board motion sensors can also be exploited to reveal any audio/video file played on the victim's smartphone loudspeaker. The attacker could exploit motion sensors, by logging the output of motion sensors during the media play, and learn about the audio played by the victim. This fact could also be exploited by advertisement agencies to spam the victim by using the information gleaned from the eavesdropped media content (e.g., favorite artist).
- *Voice Assistant Response Leakage*: Our proposed threat may extend to phone's smart voice assistant (like Google Assistant or Samsung Bixby), that communicate by reaffirming any given voice command using the phone's loudspeakers. While this action provides a better user experience, it also opens up the possibility of the attacker learning the voice assistant's responses.

Considering these attack instances, we explore the vulnerability of motion sensors to speech reverberations, from the smartphone's loudspeakers, conducted via the smartphone's body. We also examine the frequency response of the motion sensors and the hardware design of the smartphones that leads to the propagation of the speech reverberations from the phone's loudspeaker to the embedded motion sensors. Our contributions are three-fold:

(1) ***A New Speech Privacy Attack System:*** We propose a novel and lightweight attack, Spearphone (Section 4), that compromises speech privacy by exploiting the embedded motion sensor (accelerometer) of a smartphone. Our work targets speech reverberations (surface-aided and aerial vibrations), produced by the smartphone's loudspeakers, rather than the phone owner's voice, directed towards the phone's microphone. This includes privacy violation of remote caller on a voice call (live at remote end but still played through phone owner's loudspeakers), user behavior by leaking information about audio/video played on phone's loudspeakers or the smartphone's voice assistant's response to a user query (including the issued command) through the loudspeakers in a preset voice. Accelerometers are not designed to sense speech as they *passively reject air-borne vibrations* [18]. Thus, it is very hard for an attacker to eavesdrop on speech using accelerometer readings. Prior work on motion sensor exploits required the speech to be replayed via *external loudspeakers* while a *smartphone (with embedded motion sensors) was placed on the same surface as the loudspeaker.* In contrast, our work leverages the speaker inbuilt in the smartphone to provide a fundamentally different attack vector geared towards eavesdropping on *speech reverberations* (Section 2). Spearphone is a three-pronged attack that performs gender, speaker and speech classification using accelerometer's response to the speech reverberations, generated by the victim's phone's speakers.

(2) ***Attack Design and Implementation:*** As a pre-requisite to the Spearphone attack, we perform frequency response analysis of motion sensors (accelerometer and gyroscope) to determine the sensor most susceptible to our attack (Section 3). We find accelerometer to be the most receptive and therefore design our attack based on its readings associated with smartphone loudspeaker's speech signals. The attack is designed to work on the Android platform, facilitated due to the "zero-permission" nature of motion sensors (up to the latest Android 10). We execute the attack by carefully using off-the-shelf machine learning and signal processing techniques to minimize attacking efforts (Section 5). By using known techniques and tools, we believe that our attack implementation has a significant value as it can be created by low-profile attackers. Although we use standard methods to keep our attacks more accessible, we had to address several technical challenges like low sampling rates of the motion sensors and appropriate feature set selection.

(3) ***Attack Evaluation under Multiple Setups:*** We evaluate Spearphone under multiple setups mimicking near real-world usage of smartphone loudspeakers (Section 6). We show that Spearphone can perform gender and speaker classification on media playback, requiring as low as just one word of test data with an f-measure ≥0.90 and ≥0.80 respectively, which shows the threat potential of the attack. Promising, although slightly lower, classification results are obtained for the voice call and voice assistant response scenarios. The speech classification result also shows the possibility of speech identification, essentially turning it into a loudspeaker for the attacker. Our evaluation and datasets capture the three threat instances as they all require the speech signals to be output by the phone's loudspeakers.

2 BACKGROUND AND PRIOR WORK

The embedded motion sensors in smartphones could leak a user's private information by capturing the vibrations associated with users' movements such as typing on the phone's keyboard, causing sensitive information leakage on mobile devices [16, 27, 29–31]. In addition, (sp)iPhone [27] showed that the vibrations generated by typing on a physical keyboard, can be captured by a nearby smartphone's accelerometer to learn the user's input. Additionally, it is necessary to consider speech privacy in daily scenarios like private meetings, phone conversations, audio media consumption. Existing studies have shown that background noises affects MEMS sensors [17, 20, 21]. Due to the low sampling rate of motion sensors (200Hz on most smartphones), their capability of snooping on speech is often ignored. However, recent work ([11, 28, 32]) shows that the smartphone's motion sensors could reveal speech information. Specifically, [28] shows that gyroscope can measure acoustic signals from an external loudspeaker to reveal speaker information. [32] uses smartphone's accelerometer to extract signatures from the live human voice for *hotwords* extraction.

Speechless [11] analyzes the work done in [28] and defines the nature of speech propagation that could affect motion sensors. In particular, it points out that the surface propagation of speech affects the motion sensors (*surface-aided*) in [28] while aerial speech propagation (from vocal tracts to the embedded motion sensors) may lack the necessary energy to impact them. Pitchin [24] presented an eavesdropping attack using embedded motion sensors (having a higher sampling rate than a smartphone motion sensor) in an IoT infrastructure, capable of speech reconstruction.

The above studies focus on the possibility of the embedded motion sensors responding to the external sound sources (e.g., loudspeaker and live human voice). Our work explores the possibility of revealing the speech from the smartphone's built-in speakers, using the phone's own motion sensors. Compared to [28], we found that the accelerometer performs better than the gyroscope, when picking up the speech reverberations. Moreover, [28] examined the speech from an external loudspeaker, which produces stronger sound/vibration signals and only targeted the local speaker's speech using their smartphone. Smartphone loudspeakers lack the wide frequency response compared to an external loudspeaker (with woofers), especially at low frequencies. Since the speech signals producing vibrations, consist of low frequencies, our threat model is weaker than [28] and it exploits both surface-aided and aerial vibrations, propagated within the smartphone's own body. Thus, we believe [28] presents a threat model very favorable to the attacker but potentially too restrictive for the real world.

A more recent study [14] also studied accelerometer-based speech privacy inference under the setup that the accelerometer is on the same smartphone as the speaker. However, this attack relies on a sophisticated deep neural network (DNN) with fine-tuned hyperparameters, which significantly increases the attacking efforts. Moreover, this type of attack requires the training data to be collected on the victim's phone only as cross-device training/testing does not provide accurate results due to device hardware variability (even between the same model phone). Having a considerably large amount of training data, such as thousands of data samples required by [14], for the DNN training thus makes the attack less practical.

To summarize, we identify and dissect live speech and media instances in which the speech privacy attack through motion sensors works, whereas a recent study [11] concluded these sensors to be "speechless" in most other setups (humans speaking into the phone, or when the loudspeaker does not share the same surface as the phone). Our proposed setup and previous studies ([11, 28] use a similar scenario but the key difference lies in the targeted speech. In previous works, the targeted speech was from sources external to the smartphone while we consider the speech that originates from the smartphone itself via speech reverberations. Additionally, we use lightweight classifiers and an off-the-shelf machine learning tool to minimize attacking efforts and make Spearphone deployable in practice. We elaborate our detailed attack model in Section 4.

3 SENSORS VS. SPEECH REVERBERATIONS

A smartphone's body (Figure 1) provides an alternative pathway for propagating the resulting sound reverberations to the accelerometer and gyroscope in the phone, in addition to the air-borne vibrations(*air-borne propagation*). The embedded motion sensors are designed for sensing the phone motions. However, the above illustrated sound reverberations can also allow their exploitation (due to *zero-permission* nature). To measure the response of the accelerometer to the built-in loudspeaker, we play a specific signal and record the accelerometer readings with a smartphone (Samsung Galaxy Note 4). The smartphone sensor sampling rate is 250Hz and it is placed on a wood table. We generate a chirp sound signal, sweeping from frequency 0Hz to 22kHz for 5 minutes, and play it through the smartphone's built-in loudspeaker at maximum volume.

We found that the accelerometer has a strong response to the sound frequency ranging from 100Hz to 3300Hz. This is because the built-in loudspeaker and the accelerometer are on the same device, and the sound gets transmitted through the smartphone components causing vibrations. Moreover, we observed that different frequency sounds cause responses at the low frequency points of the accelerometer and generate aliased signals [28], which can be expressed by the equation $f_a = |f - N \cdot f_s|$, where f_a, f, f_s are the vibration frequency of the accelerometer, sound frequency and the accelerometer sampling rate. N can be any integer. Therefore, the accelerometer can capture rich information from the sound but with aliased signals in low frequency.

To determine the dominant propagation medium in our proposed attack, we compared the phone's accelerometer response in two settings: (1) an LG G3 phone's accelerometer captures the speech from its own loudspeaker; and (2) the G3 accelerometer captures the speech from the loudspeaker of another phone placed near it on the shared table. The volume of the played speech is adjusted at the same level, and the distance between the loudspeaker and the G3 phone's accelerometer is kept the same. Appendix Figure 4 shows the root mean square (RMS) of the captured sensor readings. We observe that our attack setting (*smartphone body*) possesses a much higher response, as high as 0.2, than the *shared solid surface* setting (around 0.05). It indicates that the smartphone body dominates the vibration propagation so as to carry more speech-relevant information in the captured accelerometer readings.

4 ATTACK OVERVIEW & THREAT MODEL

4.1 Spearphone Threat Instances

In Spearphone, we assume that the smartphone's loudspeaker is being used to output the audio. Some examples of Spearphone threat instances are described as follows:

- *Voice Call:* In this threat instance, the victim is communicating with another person and listening in the *loudspeaker mode.* We assume the phone loudspeaker is at the maximum loudness level to produce strongest speech reverberations (although we also test the effect of lower volumes and validate the threat under such conditions). The phone could be hand-held or placed on a solid surface like a table. In this threat instance, the attacker is able to capture reverberations on the victim's phone, generated in real time during the phone call.
- *Multimedia:* The live call instance could extend to situations where speech is produced by smartphone's loudspeakers while playing a media file. While the content of the media may not be private, an attacker can get some confidential information about the victim (for example, Snapchat videos, preferred music). Advertisement companies could use this information to target victims with tailor-made ads. It could be a breach of privacy if a person's habits or behavior patterns are exposed to the attacker that could potentially be used to discriminate them from jobs, insurance purposes, financial benefits, etc.
- *Assistant:* Most modern smartphones come with an inbuilt voice assistant for performing intelligent tasks. The voice assistant often confirms the user's command to ensure the desired action. It makes the process user-friendly and gives the user a choice to modify or cancel the current process. If the phone assistant uses the inbuilt phone loudspeakers, any response from the phone assistant is played back via these loudspeakers and can potentially affect the motion sensors, in turn exposing the intent of the user to an attacker exploiting the motion sensors.

4.2 Attacker's Capabilities

The attacker in our threat model has similar capabilities as elaborated in previous literature [11, 28]. The attacker can fool the victim into installing a malicious application or a malicious website to

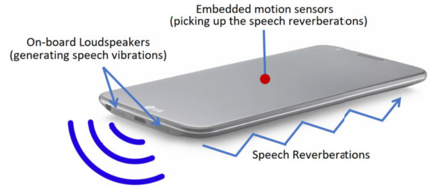

Embedded motion sensors
(picking up the speech reverberations)
On-board Loudspeakers
(generating speech vibrations)
Speech Reverberations

Figure 1: Speech reverberations, propagating within the smartphone's body, impact the motion sensors

track the motion sensor readings in the background [28]. These malicious applications could be designed to get triggered for specific threat instances and start logging the motion sensor output. For example, the attacker, using the malicious app, collects speech samples that constitute the training data, over multiple voice calls. This information could then be used to identify a speaker, in a public zoom call where speaker identification may be difficult due to multiple speakers. This threat model, is similar to some of the prior work in this line of research [14, 28].

Spearphone attempts to compromise speech privacy by performing gender, speaker, and speech classification. More specific privacy concerns for each type of classification/leakage are provided below. We also limit our threat model to utilize a finite set of words (a closed dictionary) although it could be expanded by identifying individual phonemes contained in the speech.

- *Gender Classification (Gen-Class):* Gender classification can cause a privacy compromise when the gender of a person may be used to target them. Advertising sites could push spam advertisements of products aimed towards a specific gender [19]. It can also be used for gender discrimination([19]) where job search advertisements were gender biased. Certain oppressive societies put restrictions on particular genders and may use gender classification to target individuals. Gender classification is relevant to the *Voice Call* threat instance (Section 4.1) where the attacker is interested in the identity of the remote caller that can be narrowed down by the gender of the caller. Spearphone extracts speech features from unlabeled motion sensor recordings and classifies each extracted sample as originating from either male or female speaker using classification models built on previously obtained labeled samples of sensor recordings.
- *Speaker Classification (Spk-Class):* Speaker classification involves identifying a speaker in a voice call. For example, an attacker can learn if a particular individual was in contact with the phone owner at a given time. Another example could be a person, under surveillance by law enforcement, who is in contact with the phone owner. It could also lead to leakage of the entire phone log of the phone owner. This classification is most suited for the *Voice Call* threat instance (Section 4.1) where the identity of the remote caller can directly be revealed. Similar to *Gen-Class* the attacker trains a classification model from the labeled dataset of sensor recordings, associated with a unique speaker and then tests the obtained unlabeled sensor recording against this model.
- *Speech Classification (Speech-Class):* Spearphone aims to learn the actual words during the attack. To perform *Speech-Class*, we build a classification model based on a finite word list. Speech features from the obtained sensor readings for isolated words are compared against the labeled features of the word list by the classification model that provides the attacker with a possible rendition of the actual spoken word. We also study the feasibility of performing speech reconstruction by isolating words from natural speech and then using word recognition on isolated words to reconstruct speech. Speech classification is relevant to all the three threat instances (Section 4.1). It can disclose the specifics of the call for the voice call threat scenario, leak the contents of the media consumed by the victim and reveal the actions taken by the voice assistant in response to the victim's commands.

4.3 Attack Setup

In our model, we study the speech reverberations generated from the smartphone's inbuilt speakers. Therefore, we exclude any external vibration generating source such as external loudspeakers studied in [11, 28]. Our threat model assumes the victim's phone is the only device that is present in the environment and the only vibrations present in the environment are generated by the victim's smartphone speakers. This assumption is supported by the discoveries of [11] and [18] that indicated that aerial vibrations from ambient noise are too weak to affect the MEMS accelerometers. To test the threat instances, we categorize two setups where the victim's phone speaker can impact the embedded motion sensors.

- *Surface* Setup: In this setup, the phone is kept on a flat surface with its screen facing up. This setup may be used in *Voice scenario* where the victim places the phone on a table while talking to someone with the phone on speaker mode. This setup also mimics occurrences when the phone is put on a table, countertop etc. in *Multimedia scenario* and *Assistant scenario*.This setup is similar to [28] and [11] that were primarily focused on surface-borne propagation of speech vibrations via a shared surface. However, our setup also allows the possibility of aerial propagation of speech vibrations due to the very close proximity of the speech source (the phone speaker) and the accelerometer. As both reside within the same device body, we do not rule out the effect of aerial vibrations of speech on the accelerometer and hence use the encompassing term "reverberations" as indicated in Figure 1.
- *Hand Held* Setup: The victim may also hold the phone in hand while in *Voice scenario*, playing a media file in *Multimedia scenario* or using their phone's assistant in *Assistant scenario*. In our threat model, we assume that while holding the phone in hand, the victim is stationary with no hand or body movement.

The attacker can examine the captured data in an off-line manner and use signal processing along with machine learning to extract relevant information about the intended victim.

5 ATTACK DESIGN

Spearphone relies on the smartphone loudspeakers to generate reverberations from the speech signals. We tested the ear piece speaker, that is normally used to listen to incoming phone calls (a target for our attacker). However, we did not observe any footprints of speech, indicating incapability of the ear piece speaker to produce speech reverberations strong enough to impact the accelerometer.

5.1 Motion Sensor Recording

We designed an Android application that mimics a malicious attacker (Section 4). On start, the application begins logging motion sensor readings. After a delay of five seconds, we play a single word while the application is recording motion sensor data. This step partially mimics the act of the callee's speech generated during a phone/voice call or the playing of a media file on the phone via the inbuilt loudspeakers. Our use of isolated words can also be extended to continuous speech, but we do not aim to implement a complete speech recognition system, limiting only to showcase the threat posed by embedded motion sensors. Upon completion, we process motion sensor readings as detailed in subsequent subsections.

5.2 Identifying Speech Areas

Once the attacker obtains motion sensor output from the malicious application, he needs to extract speech areas for performing *Gen-Class*, *Spk-Class* and *Speech-Class* (Section 4.2). Since we used isolated words in our attack, each speech sample contains one instance of a spoken word. As gyroscope did not display a noticeable presence of speech in the spectrum of its readings (Section 3), accelerometer is the only motion sensor that is considered in Spearphone. To extract speech from accelerometer recordings, we trim off the beginning five seconds and ending two seconds of the recordings to compensate for the initial delay before playing the isolated word and the ending finger touch for pressing the "Stop" button to pause the motion sensor recordings.

Since we see maximum response along the Z axis, for accelerometer's reaction against speech (Section 3), we try to determine the speech areas along the Z axis readings and use corresponding areas for the X and Y axes. To determine the area of speech in the Z axis readings for accelerometer, a sliding window (size=100 samples) is used. Since different words have varying lengths of utterance, we picked the duration of the shortest word as the size of sliding window. We calculate variance in each window to determine the sensor behavior within that time. A higher variance in the readings indicates the presence of an external motion (speech vibrations). We extract the bounds of window with maximum variance as the sensor readings influenced due to the presence of speech.

5.3 Feature Set for Speech Classification

Mel-Frequency Cepstral Coefficients (MFCC) are widely used in audio processing as they give a close representation of human auditory system. While MFCC features are sensitive to noise, our threat model (Section 4) assumes minimal interfering noise. Time-frequency domain features are another classification option that consist of statistical features of the signal in time domain such as minimum, maximum, median, variance, standard deviation, range, absolute mean, CV (ratio of standard deviation and mean times 100), skewness, kurtosis, first, second and third quartiles, inter quartile range, mean crossing rate, absolute area, total absolute area, and total signal magnitude averaged over time. Frequency domain features are calculated by using Fast Fourier transformation (FFT). The FFT coefficients were used to derive energy, entropy and dominant frequency ratio in time-frequency features.

We compared both MFCC and time-domain frequency features to determine the most suitable feature set for classifying the speech signals. We use the metrics (Section 5.4) and the following classifiers: Support Vector Machine (used in [28]) with Sequential Minimal Optimization (SMO), Simple Logistic, Random Forest and Random Tree (variants of the classifier used in [32]). We used the TIDigit word list [2], for using isolated words, on LG G3 in *Surface* scenario. The time-frequency features (Appendix Table 6) outperformed MFCC features, using 10-fold cross validation for all algorithms. This result (backed by [32]), led us to using it in our attack. Similarly, Random forest outperformed other classifiers using the time-frequency features (Appendix Figure 5 and Figure 6).

We further studied the distribution differences of time-frequency features for *Gen-Class*, *Spk-Class*, and *Speech-Class*. Appendix Figure 7 shows the distribution of a subset of the most salient features in box plots, which works best for *Gen-Class*. In particular, the identified feature set includes the second quartiles (Q2), third quartiles (Q3), signal dispersion (SigDisp), mean cross rate (MCR), ratio of standard deviation over mean (StdMeanR) and energy, along different axes. Similarly, we also identified the most effective time-frequency features for *Spk-Class*, and *Speech-Class* (boxplots presented in Appendix Figures 8 and 9).

5.4 Evaluation Metrics

Precision indicates the proportion of correctly identified samples to all the samples identified for that particular class. It is the ratio of number of true positives to number of elements labeled as belonging to the positive class. *Recall* is the proportion of correctly identified samples to actual number of samples of the class. It is calculated as the ratio of number of true positives to number of elements belonging to the positive class. *F-measure* is the harmonic mean of precision and recall. For perfect precision and recall, f-measure value is 1 and for worst, it is at 0.

5.5 Design Challenges

5.5.1 *Low Sampling Rates.*

Operating Systems like Android limit the data output rate for motion sensors, to conserve the battery life, valuable processing and memory power, which makes it harder to turn the on-board motion sensors into microphones. Compared to an audio microphone (sampling rate = 8kHz to 44.1kHz), motion sensors are severely limited in their sampling rate. In addition, the on-board loudspeakers may be limited in their capacity to correctly reproduce the audio. Thus, we need to choose the motion sensor that can capture most of the speech signal. We compared the frequency response of both accelerometer and gyroscope in Section 3 and the accelerometer's response was stronger than the gyroscope's response in the frequency range 100 − 3300Hz (Section 3). Thus, we make use of accelerometer in our experiments.

5.5.2 *Complete Speech Reconstruction.*

Performing speech reconstruction with the information captured by a low sampling rate and low fidelity motion sensors may not be sufficient to recognize isolated words. Moreover, it is unrealistic to generate a complete dictionary (i.e., training profile) of all the possible words for the user's full speech reconstruction. To address these issues, we extracted the time-frequency features from the accelerometer readings, which exhibit rich information to distinguish a large number of words based on existing classifiers (e.g., Random Forest and Simple Logistic). We performed word isolation by analyzing the accelerometer readings's spectrogram under natural speech and calculated the Root Mean Square of the power spectrum values. We developed a mechanism based on searching the keywords (e.g., credit card number, targeted person's name and SSN) and only used a small-sized training set to reveal more sensitive information while ignoring the propositions, link verbs and other less important words.

6 ATTACK EVALUATION

6.1 Experiment Setup

Smartphones: We conducted our experiments using four smartphones: LG G3, Samsung Galaxy S6, Galaxy S8 and Note 4. The experiments were performed in a quiet laboratory on a hardwood table-top for *Surface* setup, while the *Hand Held* setup was conducted by two participants holding the phone in their hands.

- *Operating System:* We focused on Android based smartphones as they do not require explicit user permission to obtain access to motion sensor data. In contrast, the iOS mobile operating system (version 10.0+) requires any application accessing motion sensor data to state its intent in the key "NSMotionUsageDescription". The intent would be displayed to the user and failure to state its intent results in immediate application exit. As pointed out in Section 1, the sizeable market share of Android (worldwide and the US) allows us to treat the threat posed to smartphones operating on this platform with extreme concern.
- *Sensors:* The accelerometer embedded in the smartphones in our experiments had an output data rate of 4-4000 Hz and an acceleration range of ±2/±4/±8/±16g. The liner acceleration sensitivity range are 0.06/0.12/0.24/0.48 mg/LSB. A comparison with the LSM6DSL chip used in the latest Samsung Galaxy S10 smartphone indicates similar properties for the accelerometer.

Word Datasets: We used the subset of TIDigits corpus ([2]). It contains 10 single digit pronunciation from "0" to "9" and 1 additional pronunciation "oh". It contains 5 male and 5 female speakers, pronouncing the words twice. The sampling rate for the audio samples is 8kHz. We also used a pre-compiled word list (PGP words Dataset) uttered by Amazon Mechanical Turk workers in a natural environment. The list consisted of fifty-eight words from PGP words list and they were instructed to record the words in a quiet environment. This data collection activity was approved by the university's IRB and the participants had the choice to withdraw from the experiment at any given time. We used 4 male and 4 female Amazon Turk workers' audio samples (44.1 kHz sampling frequency). PGP word list is used for clear communication over a voice channel and is predominantly used in secure VoIP applications.

Speech Processing: We used Matlab for processing the accelerometer output performing feature extraction (Section 5). We used Weka [12] to perform gender, speaker and speech classification on the extracted speech features. In particular, we test the attack with Random Forest classifier that outformed other classifiers as noted in Section 5.5. We used default parameters for the classification algorithm and the detailed configurations are listed in Appendix Table 5. We used both 10-fold cross-validation and the training and testing methods for classification. 10-fold cross-validation partitions the sample space randomly in 10 disjoint subspaces of equal size, using 9 subspaces as training data and retaining 1 subspace as testing data. For training and testing method, we split the dataset into training set and test set with the split being 66% of the dataset being used for training and remaining 34% being used for testing.

In our attack, the attacker collects the training samples for building the classifier, which is unique for each device. Since our dataset is not large (limited to 58 words for PGP words and 22 words for TIDigits), we believe that it does not indicate a significant overhead for the attacker to procure the training samples for each device targeted under the attack. Most other motion sensor attacks to our knowledge (e.g., [16, 27, 27, 29–31]), including Gyrophone, have similar or even more strict training requirements for the attacker.

Table 1: Gender and speaker classification (10 speakers) for Surface setup using TIDigits and PGP words dataset using Random Forest classifier and time-frequency features

	10-fold cross validation		Test and train	
	TIDigits	**PGP words**	**TIDigits**	**PGP words**
Gender classification				
Samsung Galaxy S8	0.98	0.99	0.97	0.99
Samsung Galaxy S6	0.91	0.80	0.87	0.82
Samsung Note 4	0.99	0.91	1.00	0.95
LG G3	0.89	0.95	0.85	0.95
Speaker classification				
Samsung Galaxy S8	0.88	0.90	0.89	0.93
Samsung Galaxy S6	0.69	0.70	0.56	0.71
Samsung Note 4	0.94	0.80	0.92	0.80
LG G3	0.91	0.92	0.89	0.95

Effect of Noise: In our threat model, the loudspeaker resides on the same device as the motion sensors. Thus, any reverberations caused by the device's loudspeaker would impact the motion sensors. [11] and [18] claimed that external noise in human speech frequency range, traveling over the air, does not impact the accelerometer. Ba et al. [14] concluded that airborne acoustic noises at regular frequency (below 22000Hz) and sound pressure level are unlikely to distort the accelerometer measurements. Hence, any such noise in the surroundings of the smartphone would be unable to affect the accelerometer's readings. The speech dataset used in our experiment, *PGP words dataset*, was collected from Amazon Mechanical Turk workers, recording their speech in environments with varying degree of background noise. This dataset thus imitates the speech samples that the attacker may face in the real-world, such as during our attack instances involving phone calls.

6.2 Gender and Speaker Classification in Voice Call Instance (*Surface* Setup)

6.2.1 Surface Setup using TIDigits. The results for the *Surface* setup, where the victim's phone is placed on a surface such as a table, using TIDigits dataset is shown in Table 1 for *Gen-Class* and *Spk-Class*. We observe that the attack was able to perform *Gen-Class* with an f-measure > 0.80 with the attack being particularly successful on Galaxy S8 and Note 4 as demonstrated in Table 1. As a baseline, the scores are significantly better than a random guess attacker (0.50) indicating the success of the attack in this setup. For *Spk-Class*, we note that the attack is more successful on Samsung Galaxy S8, LG G3 and Note 4 when compared to Galaxy S6 with f-measure > 0.60. A random guess attack performance is significantly worse at 0.10 (for 10 speakers) when compared to this attack.

6.2.2 Surface Setup using PGP words dataset. In Table 1 for *Gen-Class* and *Spk-Class*, comparing the attack against a random guess attack (0.50), we observe that the reported f-measure for the attack on all phone models was more than 0.70 in both 10-fold cross-validation and train-test model. The attack on LG G3 and Samsung S8 had an f-measure of over 0.90 consistently across all the tested classification algorithms. Table 1 show Spearphone's performance when *Spk-Class* was performed using the PGP words dataset.

For a 10-speaker classification model, a random guess attack would give us an accuracy of 0.10. In our tested setup, we were able to achieve much higher f-measure scores with the attack on LG

Table 2: Gender and speaker classification (10 speakers) for Hand Held setup using TIDigits and PGP words dataset using Random Forest classifier and time-frequency features

	10-fold cross validation		Test and train	
	TIDigits	PGP words	TIDigits	PGP words
Gender classification				
Samsung Galaxy S6	0.77	0.72	0.76	0.70
Samsung Note 4	0.81	0.87	0.77	0.88
LG G3	0.99	0.95	1.00	0.95
Speaker classification				
Samsung Galaxy S6	0.33	0.34	0.26	0.29
Samsung Note 4	0.73	0.75	0.61	0.70
LG G3	0.98	0.93	1.00	0.95

G3 and Samsung S8 achieving a score of almost 0.90. The attack on Galaxy S6 performed the worst among all the phones but still had a better f-measure score of over 0.50 when compared to the baseline random guess attack. These results lead to conclusion that Spearphone threat could be performed using *Spk-Class* in this setup.

We also performed the binary classification for speakers by using two classes "Targeted Speaker" and "Other", that categorizes each data sample as either in the voice of the target speaker or any other speakers. We used PGP words dataset in our evaluation as it contained more words per speaker compared to TIDigits dataset. Using Random Forest classifier and 10-fold cross-validation, the mean f-score for this binary speaker classification for LG G3 was 0.97, for Galaxy S6 was 0.90, and for Note 4 was 0.94.

6.3 Gender and Speaker Classification in Voice Call Instance (*Hand Held* Setup)

6.3.1 Hand-held Setup using TIDigits dataset. In Table 2 for *Gen-Class*, we observe that the performance of the attack on LG G3 is better when compared to other devices for both 10-fold cross-validation model and train-test model with overall f-measure being approximately 0.70, which is better than a random guess attacker (0.50). For *Spk-Class*, we see that the scores of Galaxy S6 are worse when compared to LG G3 with Note 4 having scores in between these devices. The f-measure values for LG G3 for *Spk-Class* are over 0.90 for all the tested classifiers, for Note 4 these values are over 0.50 while Galaxy S6 values hover around 0.25. When compared to a random guess attack (0.10), the attack on G3 is better while on Galaxy S6 it is slightly better.

6.3.2 Hand-held Setup using PGP words dataset. The *Gen-Class* attack result is shown in Table 2. The 10-fold cross-validation model indicates that the f-measure value of the attacker's classifier for LG G3 is the best performer among all three phone models. Similar to *Surface*, the attack performed better than a random guessing attacker (0.50) while the performance of attack was similar to the performance in *Surface* setup. The attack's evaluation for *Spk-Class* (Table 2) shows that the attack is able to perform speaker identification for LG G3. The f-measure values, however, drop for Note 4 while the performance is worst for Galaxy S6. Thus, the attack's performance, while still better than a random guess attack (0.10), suffers a bit of setback for Note 4 and more so for Galaxy S6. The binary classification for speakers (previously described in *Surface* setup) shows that the f-measure values when the smartphone is hand-held (*Hand Held* setup) are similar to *Surface* setup. The f-measure score averaged for 8 speakers with LG G3 was 0.97, for Galaxy S6 was 0.84, and for note 4 was 0.92.

Table 3: Effect of loudness on gender and speaker classification accuracy using Samsung Note 4 for Surface setup using Random Forest classifier and time-frequency features.

		Volume Level		
		75%Vol_{max}	80%Vol_{max}	Vol_{max}
Gender Classification	*TIDigits*	0.93	0.90	0.99
	PGP word	0.78	0.95	0.91
Speaker Classification	*TIDigits*	0.45	0.70	0.94
	PGP words	0.54	0.79	0.80

6.4 Result Summary and Insights

The speaker classification accuracies for Note 4 and Galaxy S6 are higher for PGP words compared to TIDigits. This may be because PGP words dataset (sampled at 44.1kHz) was recorded at a higher sampling rate when compared to TIDigits (8kHz). This effect is not prominent in LG G3 because the sampling rate of its motion sensors is slightly lower (120Hz) than Note 4 or Galaxy S6 (around 200Hz) or S8(around 400 Hz). Ba et al. [14] performed a scalability study using Samsung Galaxy S8 (420 Hz), Huawei Mate 20 (500 Hz) and Oppo R17 (410Hz) and found an increase in the classification model accuracy. Our results with Galaxy S8 in Table 1 are in-line with the scalibility study, showing a better performance when compared to other smartphones with lower sampling rates. The gender and speaker classification accuracies seem to decrease a bit for the PGP words dataset in some instances. We believe that due to some background noise present in PGP words dataset, the accuracies may have been affected negatively. The accuracies of LG G3 do not seem to be impacted though, which we believe maybe due to its lower sampling rate (making it less prone to data degradation).

Effect of Surface: Another interesting observation is that the *Surface* setup overall produces better classification results than the *Hand Held* setup. The hand motions and body movements are negative influences, but they only cause low frequency vibrations, which have been removed by our high-pass filter. Another possible explanation could be the vibration absorption/dampening caused by the holding hand. To test this reasoning, we conducted experiments with the Note 4 phone placed on a soft surface (i.e., soft couch). The gender classification accuracy is 87.5%, similar to the handheld scenario (87%), both of which are lower than the hard tabletop scenario. This suggests vibrations are possibly being absorbed by the hand to some degree. The speaker classification results overall seem similar to speaker classification using audio recordings [25]. This behavior may be an indication that prominent speech features present in audio vibrations are also picked up by the accelerometer, as showcased by our experiments. Comparing our results with [28], we find that they achieved the best case gender classification accuracy of 84% using DTW classifier on Nexus 4, which is lower than our accuracy of almost 100% using Random Forest classifier on Samsung Note 4, using the same dataset (TIDigits). For speaker classification, we obtained a higher accuracy of over 90% using Random Forest classifier on Samsung Note 4 while that for Micalevsky et al. [28] was only 50% for mixed gender speakers using DTW classifier for the same dataset (TIDigits). There is still room for improving the accuracy by exploring more features and deep learning methods (similar to [14]), which will be explored in our future work.

Effect of Loudness: We also evaluate the impact of the smartphone speaker volume on the performance of Spearphone. We test the gender and speaker classification performance of Spearphone when setting the smartphone speaker volume to 100%, 80%, and 75%

of the maximum volume. Table 3 presents the results for Samsung Note 4 phone, when it is placed on the table (i.e., *Surface* setup). The results show that while lower volume does impact the accuracy negatively, the lower volumes still achieve a robust accuracy (i.e., 80% volume achieves 95% accuracy for gender classification and 79% accuracy for speaker classification with the PGP words dataset). Also, the results indicate that the lower volume still causes privacy leakage, when compared to the random guessing accuracy (i.e., 50% for gender classification and 10% for speaker classification).

People tend to use maximum volume to make the speech clear and comprehensible to avoid missing any important information [8]. The louder volume, while providing clearer speech, would expose speech privacy more significantly via our Spearphone attack. In addition, we believe that the quality of the speakerphones on smartphones will improve over time and there are also powerful speaker cases in use today that can be physically attached to the phones [6, 7], and speech leakage over such higher quality speakerphones could be more plausible, even at lower volume levels.

Natural Speech Dataset: While Spearphone achieves a high accuracy for the isolated word data set (i.e., TIDigits/PGP words), we further evaluated the performance of Spearphone with a natural speech dataset (VoxForge [1]), which provides samples of sentences (10 words long on average) spoken by 5 male and 5 female speakers, with 100 samples for each speaker. In particular, for speaker classification, Spearphone achieves 91.3% with LG G3 using Random Forest for 10-speaker classification under 10-fold cross validation. The result is very similar to the speaker classification with the isolated word datasets, which indicates that the possibility of the attack in a practical natural speech scenario.

Realistic Voice Call Scenario: To evaluate the threat of Spearphone in more realistic scenarios like a real voice call, we downgraded the sampling rate of our PGP words dataset to 8kHz. The gender and speaker classification results (using f-measure scores) on Samsung Note 4 for the dataset using random forest classifier and 10-fold cross validation method were 0.73 and 0.47. For LG G3, the gender and speaker classification results measured as f-measure were 0.99 and 0.60. Compared to Table 1, we see an expected drop in the speaker classification accuracy for the down-sampled PGP words dataset. The gender classification accuracy degrades for Note 4 but such opposite behavior is observed for LG G3.

6.5 Speech Recognition in Voice Calls

We next demonstrate the feasibility of speech recognition using Spearphone. We found that the G3 phone on a wooden table surface exhibited better performance, when revealing speaker information. Towards this end, we utilized G3 on a wooden table to investigate the feasibility of *Speech-Class*. We compared the performance of using time-frequency features with that of MFCC features, and found that time-frequency features give better classification accuracy than MFCC features. We also noted that random forest classifier outperformed the other tested classifiers, so we used Random Forest as our classifier on time-frequency features.

6.5.1 Speech-Class for Single Speaker. **TIDigits dataset:** Table 4 shows Spearphone's accuracy of recognizing a single speaker's 11 isolated digit numbers (TIDigits dataset). For 10-fold cross validation, using time-frequency features, we achieved an f-measure of 0.74 with Random Forest classifier. In comparison, a random guess attacker would achieve an accuracy of 0.09. Similar results were obtained using train-test method for classification (Table 4), though there was a slight decrease in recognition accuracy.

Table 4: Speech recognition results for PGP words and TIDigits datasets using Random Forest classifier and time-frequency features on LG G3

	10-fold cross validation		Test and train	
	TIDigits	PGP words	TIDigits	PGP words
Single Speaker	0.74	0.81	0.62	0.74
Multiple speakers	0.80	0.75	0.71	0.67

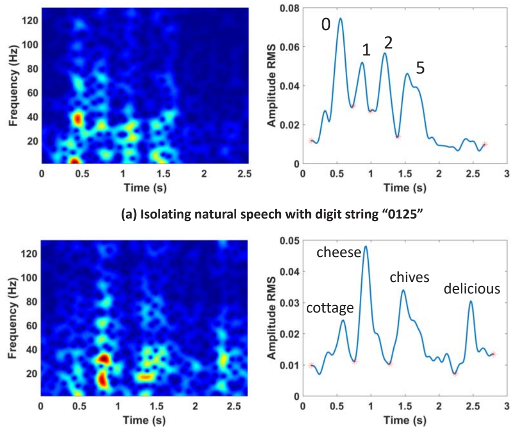
PGP words dataset: We further experimented with PGP words to explore how accurate Spearphone could recognize the isolated words other than the digits. Table 4 shows the *Speech-Class* results under 10-fold cross validation. By using the time-frequency features, Spearphone achieved a much higher f-measure score of 0.81 in recognizing words in a 58-word list than digits. In comparison, the random guess accuracy was only 0.02 for the dataset. The results of the train-test model showed a slight decrease in performance.

6.5.2 Speech-Class for Multiple Speakers. There are plenty of scenarios involving multiple people's voices presenting on a single phone such as conference calls via Skype. We studied the feasibility of speech recognition from multiple speakers. In particular, we involve two speakers (one male; one female). Table 4 also shows the f-measure scores when recognizing digit numbers from the two speakers (multiple speaker scenario). We got an f-measure score of 0.80 for the TIDigits dataset while the f-measure score for PGP words dataset, for multiple speaker scenario, was 0.75. We also used the PGP dataset, downsampled to 8kHz, to mimic real world telephony voice quality. The speech recognition accuracy for multiple speakers was 0.61, which as expected, is lower than the original dataset but still above the random guess accuracy (i.e., 0.017).

Gyrophone [28] also carried out the speech recognition task by using TIDigits dataset and 44 recorded words. However, they addressed a totally different attack setup where the sound sources were from an external loudspeaker and can achieve an accuracy of up to 0.65. Our results of speech recognition accuracy around 0.82 indicate the vulnerability of smartphone's motion sensors to its own loudspeaker's speech. Using the speech recognition and speaker identification, Spearphone is capable of associating each recognized word to the speaker identity in multi-speaker scenarios.

6.6 Speech Recognition in Multimedia and Voice Assistant Instances

We also evaluated the Spearphone accuracy in multimedia and voice assistant threat instances (Section 4.1). We used the same techniques that we used for *Speech-Class* in voice calls (section 6.5. We simulated the multimedia threat instance by utilizing VoxCeleb dataset. VoxCeleb dataset [9] is a large-scale audio-visual dataset of human speech, extracted from interview videos of celebrities uploaded to YouTube. We used a single speaker, 100 word dataset (where word truncation was done manually to extract each word) and the average length of a word in the 100 word dataset was 7.2 characters. Using random forest classifier on time-frequency features and 10-fold cross validation method, we were able to achieve a speech recognition accuracy of 0.35 for LG G3. The classification accuracy when the dataset was reduced to 58 words (for comparison



(a) Isolating natural speech with digit string "0125"

(b) Isolating natural speech with sentence "Cottage cheese with chives is delicious"

Figure 2: Word isolation using the RMS of the accelerometer spectrum

with the PGP word dataset performance) was 0.54 for LG G3. Compared to the speech classification accuracy for PGP words dataset in Table 4 for LG G3, we see a decrease in the accuracy from 0.81 to 0.54. A random guess attack has an accuracy of 0.01 (for 100 words) and 0.017 (for 58 words) indicating that our attack outperforms it by an order of 30. We attribute the decrease in the classification accuracy to the existing noise in the Youtube recordings.

We used Alexa voice assistant and generated PGP words dataset in Alexa's voice using the text-to-voice tool [5]. The text-to-voice tool pairs with the Alexa voice assistant and provides a text input feature to the user that is redirected to Alexa for repeating the user input. The speech recognition accuracy for the 58 words PGP word dataset was 0.31 for LG G3. This classification accuracy is again lower than the one reported in Table 4 for LG G3 on PGP words dataset, which was 0.81. Compared to a random guess attack, the proposed attack outperforms it by a magnitude of 18. However, Alexa's voice assistant is not human voice, albeit an artificially generated voice. Our feature set described in Section 5.3 was tuned for recognizing characteristics of reverberations resulting from human voices. We propose reevaluating the feature set in our future work that is tuned based on artificially generated voices.

6.7 Speech Reconstruction (Natural Speech)

We have shown the capability of Spearphone to recognize isolated words with high accuracy. To reconstruct natural speech, Spearphone performs *Word Isolation* and *Key Word Search*, which first isolates each single word from the sequence of motion sensor readings and then searches for sensitive numbers/words from isolated words based on speech recognition introduced in Section 6.5.

6.7.1 Word Isolation. In order to reconstruct natural speech, the words of the speech need to be first isolated from the motion sensor readings and then recognized individually. However, isolating the words from the low sampling rate and low fidelity motion sensor readings is hard. To address this challenge, we calculated the Root Mean Square (RMS) of the motion sensor's spectrum at each time point and then located local peaks based on a pre-defined threshold to isolate each word. Figure 2 illustrates an example of isolating a TIDigit string ("0125") and a PGP sentence ("Cottage cheese with chives is delicious"). The motion sensor's spectrograms were converted to the amplitude RMSs at the right side of the figure. Based on the derived amplitude RMS, the valleys between the local peaks were detected to segment the critical words. We observed that some propositions and link verbs (e.g., "with" and "is") could hardly be detected, but this drawback has minimal effect on our results as these words do not affect the ability to understand an entire sentence. We further evaluated our word isolation method by testing 20 sentences containing around 28 words per sentence, and achieved 82% isolation success rate. By excluding the less important propositions and link verbs, we achieve around 96% success rate.

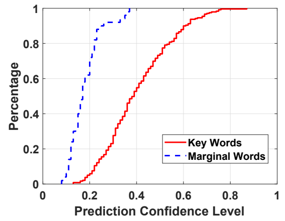


Figure 3: CDF of the prediction confidence level

6.7.2 Key Word Search. Key word search is also significant when addressing natural speech. As it is hard to train all the potential words of a natural speech beforehand, the adversary might be more interested in the sensitive numbers/words (*key words*) (e.g., credit card information, an important person's name, SSN, etc.) while marginal words such as propositions, link verbs and other such words can be ignored. Thus, a limited-size dataset is sufficient for stealing sensitive information.

After obtaining the isolated words, an adversary could search for key words based on a pre-constructed training model. In particular, Spearphone relies on the predication probability returned by the training model as the confidence level to filter the key word search results. Figure 3 shows the CDF of prediction confidence levels when 2/3 PGP words are used as key words. We observed that the key words have higher confidence levels compared to marginal words. Thus, we could apply a threshold-based method to only focus on recognizing the keywords. Further combination of word isolation and key word search to reconstruct natural speech, requires fine-grained segmentation of the words and usage of Hidden Markov/other linguistic models for word corrections. This work is an avenue for possible future work.

7 DISCUSSION AND FUTURE WORK

Attack Limitations: In our experiments, we initially put the smartphone loudspeakers at maximum volume to produce the strongest reverberations in the body of the smartphone, for maximum impact on the accelerometer. In reality, the loudness of different phones model varies and is selective per user. Hence, we tested the effect of loudness on the attack's accuracy and found out that decreasing the volume from maximum to 80%, still allowed the gender and speaker classification (though lower than the full-volume attack).

While our experiments tested two different datasets, they are still limited to single word pronunciations and are limited in size. However, single word accuracy can be extended to full sentence reconstruction using language modeling techniques. Moreover, TIDigits dataset, while relatively small, can still be effective in identifying sensitive information that mainly consists of digits. Personal information such as social security number, birthday, age, credit card details, banking account details etc., consist mostly of numerical digits. So, we believe that the limitation of our dataset size should not significantly downplay the perceived threat level of our attack.

Another factor to consider is the hand movement of the victim while holding the smartphone. Our attack experiment involved placing the phone either on a surface or held stationary in hand. Both these setups keep the smartphone stationary. Accelword [32] analyzed the impact of hand/body movements on accelerometers embedded in the smartphones and concluded that a cutoff frequency of 2 Hz would filter out the effect of these motions. Application of such a filter could make the proposed attack compatible with mobile setups, where the smartphone is not stationary.

Impact of Hardware Design: The speaker and the accelerometer specifications are different across various smartphone models. The accelerometers of the three models are similar but the loudspeaker of Galaxy S6 is less powerful, which may account for lower accuracy results on S6, especially in *Hand Held* where there is no contact between the smartphone's body and a solid surface. Besides, the positions of the speaker and the accelerometer on the smartphone may cause the acceleration patterns to respond to the same speech word differently. This is because the reverberations caused by the sound may transmit through different routes and get affected by different complex hardware components. Appendix Figure 10 shows the motion sensor specifications for some popular brands of smartphones [1]. For example, the speakers of LG G3 and Note 4 are at the back of the smartphone, which can generate different levels of reverberations when placed on the table. In comparison, Galaxy S6's speaker is located at the bottom edge of its body, thereby having a diminished effect when placed on the table.

Accelerometer Models: The three phone models tested in this paper are embedded with the Invensense accelerometer. We further analyzed the frequency response of another smartphone (Samsung Galaxy S3 having the STMicroelectronics accelerometer chip), to speech signals played via onboard loudspeaker. Our analysis suggests that the response is similar to the LG G3 (Invensense accelerometer) and both accelerometers show the frequency range between 300Hz and 2900Hz. With the MEMS technology getting better and the loudspeakers being louder and more refined with every new generation of smartphone, we believe our attack should raise more concerns about speech privacy from this perspective.

Potential Countermeasures: Most side channel attacks exploiting motion sensors center around the *zero permission* nature of these sensors. Android platform could implement stricter access control policies to restrict the usage of these sensors. Users should also be made aware of the implications of the permissions granted to the applications. However, a stricter access control policy could affect the usability of the smartphones. Even explicit usage permission model often does not work, since users do not pay proper attention to the asked permissions [23]. Moreover, many apps are designed to be overprivileged by developers [22]. Another countermeasure could filter sensitive speech frequencies from the captured readings. However, due to signal aliasing, vibrations of a wide range of frequencies are mapped non-linearly to the low sampling rate accelerometer data. Both the higher frequencies and lower frequencies contain the speech information. Thus, simply applying filters to remove the upper or lower frequencies cannot mitigate this attack.

A potential defense against Spearphone could also be set up by altering the hardware design of the phone. The motion sensors could be insulated from the the phone's speakers vibrations. To mask or dampen the vibrations leaked from the phone's speakers, surrounding the speakers with vibration dampening material may work. Speaker isolation pads are already in use in recording studios for limiting sound vibration leakage [3]. Other solutions like [13] also exist that dampen the surface-aided vibration propagation and may be useful in preventing leakage of speech vibrations within the smartphone. Further work is necessary to evaluate such a defensive measure against the threat studied in the paper.

Comparison with Previous Works: Michalevsky et al. [28] using the *gyroscope* sensor on the smartphone, achieved gender and speaker identification rate of 84% and 50% for a set of 10 speakers, using TIDigits (a small dictionary containing only digits pronunciations). Our results on TIDigits and an additional PGP words dictionary, present an improved gender and speaker identification rate of over 90% on a multitude of smartphones, using the *acceleroemter* and the speech sensing smartphone motion sensor. [11, 15] previously indicated that the accelerometer may be more responsive to speech vibrations. We also utilize this fact in our experiments to achieve a better recognition rate. Ba et al. [14] proposed a similar work using deep learning neural networks, to achieve a speaker recognition accuracy of 70% for 20 speakers. When compared, our work achieves a higher recognition rate, using a lightweight random forest classifier, albeit with a different word dictionary.

8 CONCLUSION

We proposed a novel side-channel attack that compromises the phone's loudspeaker privacy by exploiting accelerometer's output impacted by the emitted speech. This attack can leak information about the remote human speaker (in a voice call) and the speech that is produced by the phone's speaker. In the proposed attack, we use off-the-shelf machine learning and signal processing techniques to analyze the impact of speech on accelerometer data and perform gender, speaker and speech classification with a high accuracy.

Our attack exposes a vulnerable threat scenario for accelerometer that originates from a seemingly inconspicuous source (phone's inbuilt speakers). This threat can encompass several usage instances from daily activities like regular audio call, phone-based conference bridge inside private rooms, hands-free call mode and voicemail/messages played on the phone. This attack can also be used to determine a victim's personal details by exploiting the voice assistant's responses. We also discussed some possible mitigation techniques that may help prevent such attacks.

9 ACKNOWLEDGEMENT

We would like to thank our shepherd, Dr. Naofumi Homma, and the anonymous reviewers for their insightful comments and constructive feedback on the paper.

REFERENCES

[1] [n.d.]. Voxforge. http://www.voxforge.org/.
[2] 2016. Clean Digits. http://www.ee.columbia.edu/~dpwe/sounds/tidigits/.

[1] https://www.gsmarena.com/

[3] 2016. Practical sound & vibration proofing with speaker isolation pads. http://www.andrehvac.com/blog/vibration-control-products/practical-sound-vibration-proofing-speaker-isolation-pads/.

[4] 2018. Mobile Operating System Market Share Worldwide (Dec 2017-Dec 2018). http://gs.statcounter.com/os-market-share/mobile/.

[5] 2019. ASK ALEXA TO SAY WHATEVER YOU WANT. https://texttovoice.io/.

[6] 2019. big sound in a snap. https://goo.gl/k8epdz.

[7] 2019. PolarPro Beat Pulsar. https://goo.gl/PzuZFP.

[8] 2019. Volume Booster and Audio Enhancement Tips for Smartphones and Tablets. https://www.lifewire.com/boost-volume-on-phone-and-tablet-4142971.

[9] 2019. VoxCeleb: A large scale audio-visual dataset of human speech. http://www.robots.ox.ac.uk/~vgg/data/voxceleb/.

[10] Ahmed Al-Haiqi, Mahamod Ismail, and et al. 2013. On the best sensor for keystrokes inference attack on android. *Procedia Technology* 8 (2013), 947–953.

[11] S Abhishek Anand and Nitesh Saxena. [n.d.]. Speechless: Analyzing the Threat to Speech Privacy from Smartphone Motion Sensors. In *IEEE S&P 2018*. 116–133.

[12] Machine Learning Group at the University of Waikato. 2017. Weka 3: Data Mining Software in Java. https://www.cs.waikato.ac.nz/ml/weka/index.html.

[13] GC Audio. [n.d.]. VIBRATION: ORIGINS, EFFECTS, SOLUTIONS. https://www.gcaudio.com/tips-tricks/vibration-origins-effects-solutions/.

[14] Zhongjie Ba, Tianhang Zheng, Xinyu Zhang, Zhan Qin, Baochun Li, Xue Liu, and Kui Ren. 2020. Learning-based practical smartphone eavesdropping with built-in accelerometer. In *Proceedings of the Network and Distributed Systems Security (NDSS) Symposium*. 23–26.

[15] David Berend, Shivam Bhasin, and Bernhard Jungk. 2018. There Goes Your PIN: Exploiting Smartphone Sensor Fusion Under Single and Cross User Setting *(ARES 2018)*. Article 54.

[16] Liang Cai and Hao Chen. 2011. TouchLogger: Inferring Keystrokes on Touch Screen from Smartphone Motion. *HotSec* 11 (2011), 9–9.

[17] Simon Castro, Robert Dean, Grant Roth, George T Flowers, and Brian Grantham. [n.d.]. Influence of acoustic noise on the dynamic performance of MEMS gyroscopes. In *IMECE 2007*. 1825–1831.

[18] Robert F. Coleman. 1988. Comparison of microphone and neck-mounted accelerometer monitoring of the performing voice. *Journal of Voice* (1988).

[19] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated Experiments on Ad Privacy Settings: A Tale of Opacity. In *Choice, and Discrimination", Proceedings on Privacy Enhancing Technologies*. 92–112.

[20] Robert N Dean and et al. 2007. On the degradation of MEMS gyroscope performance in the presence of high power acoustic noise. In *IEEE ISIE 2007*. 1435–1440.

[21] Robert Neal Dean and et al. 2011. A characterization of the performance of a MEMS gyroscope in acoustically harsh environments. *IEEE Transactions on Industrial Electronics* (2011), 2591–2596.

[22] Adrienne Porter Felt, Erika Chin, Steve Hanna, Dawn Song, and David Wagner. 2011. Android Permissions Demystified. In *ACM CCS 2011*. 627–638.

[23] Adrienne Porter Felt and et al. [n.d.]. Android Permissions: User Attention, Comprehension, and Behavior. In *SOUPS 2012*. 3:1–3:14.

[24] J. Han and et al. [n.d.]. PitchIn: Eavesdropping via Intelligible Speech Reconstruction Using Non-acoustic Sensor Fusion. In *ACM/IEEE IPSN 2017*.

[25] E. Khoury and et al. 2013. The 2013 speaker recognition evaluation in mobile environment. In *2013 International Conference on Biometrics (ICB)*. 1–8.

[26] J. Mantyjarvi, M. Lindholm, and et al. 2005. Identifying users of portable devices from gait pattern with accelerometers. In *IEEE ICASSP*. ii/973–ii/976.

[27] Philip Marquardt and et al. 2011. (sp) iphone: Decoding vibrations from nearby keyboards using mobile phone accelerometers. In *ACM CCS 2011*. 551–562.

[28] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. 2014. Gyrophone: Recognizing Speech from Gyroscope Signals.. In *USENIX Security Symposium*. 1053–1067.

[29] E. Miluzzo, A. Varshavsky, and S. Balakrishnan. 2012. TapPrints: Your Finger Taps Have Fingerprints. In *Proceedings of ACM MobiSys*.

[30] Emmanuel Owusu and et al. [n.d.]. ACCessory: password inference using accelerometers on smartphones. In *HotMobile 2012*.

[31] Zhi Xu, Kun Bai, and Sencun Zhu. [n.d.]. Taplogger: Inferring user inputs on smartphone touchscreens using on-board motion sensors. In *ACM WiSec 2012*. 113–124.

[32] Li Zhang, Parth H Pathak, Muchen Wu, Yixin Zhao, and Prasant Mohapatra. 2015. Accelword: Energy efficient hotword detection through accelerometer. In *ACM MobiSys 2015*. 301–315.

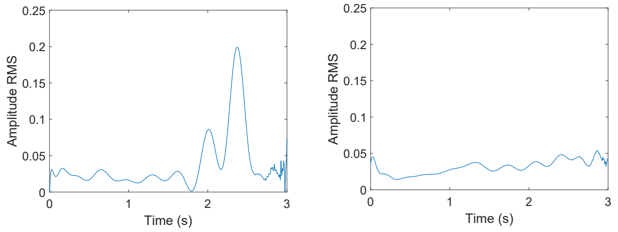
A APPENDIX

A.1 Classifier Configurations

Table 5: Configurations of tested classifiers

Classifier	Configurations
SimpleLogistic	-I 0 -M 500 -H 50 -W 0.0
SMO	-C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K -kernal PolyKernel -E 1.0 -C 250007 -calibrator Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
RandomForest	-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
RandomTree	-K 0 -M 1.0 -V 0.001 -S 1

A.2 Accelerometer Response with Different Propagation Medium

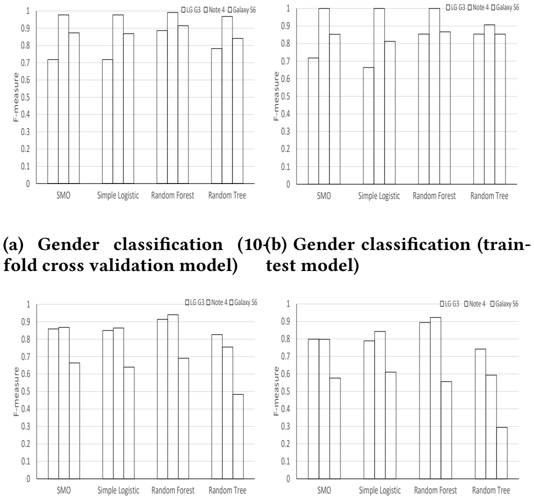


(a) Smartphone body

(b) Shared solid surface

Figure 4: The RMS of the accelerometer's response to the two experimental settings: (1) Smartphone body: the phone's accelerometer captures the reverberations from the phone's own loudspeaker; and (2) Shared solid surface: the phone's accelerometer captures the vibrations from another phone's loudspeaker via the shared solid surface.

A.3 Comparison of Various Classifiers



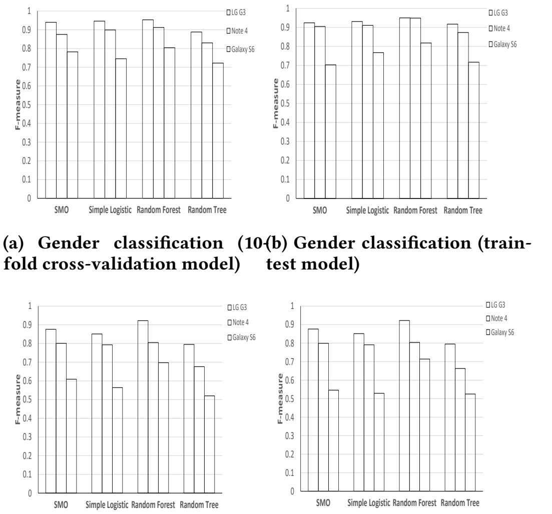
(a) Gender classification (10-fold cross validation model)

(b) Gender classification (train-test model)

(c) Speaker classification (10-fold cross validation model)

(d) Speaker classification (train-test model)

Figure 5: Gender and speaker classification (10 speakers) for Surface setup using TIDigits dataset



(a) Gender classification (10-fold cross-validation model)

(b) Gender classification (train-test model)

(c) Speaker classification (10-fold cross-validation model)

(d) Speaker classification (train-test model)

Figure 6: Gender and speaker classification (10 speakers) for Surface setup using PGP words dataset

A.4 Time-frequency Feature List

Table 6: The time-frequency features calculated from accelerometer readings of X, Y and Z axis over a sliding window

Time Domain
Minimum; Maximum; Median; Variance; Standard deviation; Range
CV: ratio of standard deviation and mean times 100
Skewness (3rd moment); Kurtosis (4th moment)
Q1, Q2, Q3: first, second and third quartiles
Inter Quartile Range: difference between the Q3 and Q1
Mean Crossing Rate: measures the number of times the signal crosses the mean value
Absolute Area: the area under the absolute values of accelerometer signal
Total Absolute Area: sum of Absolute Area of all three axis
Total Strength: the signal magnitude of all accelerometer signal of three axis averaged of all three axis
Frequency Domain
Energy
Power Spectral Entropy
Frequency Ratio: ratio of highest magnitude FFT coefficient to sum of magnitude of all FFT coefficients

A.5 Salient Features for Gender, Speaker, and Word Classification

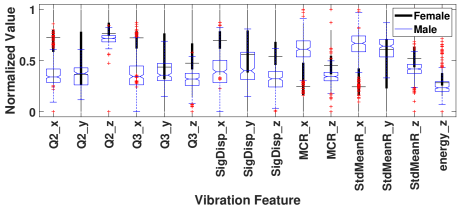


Figure 7: Salient time-frequency feature distributions for Gen-Class.

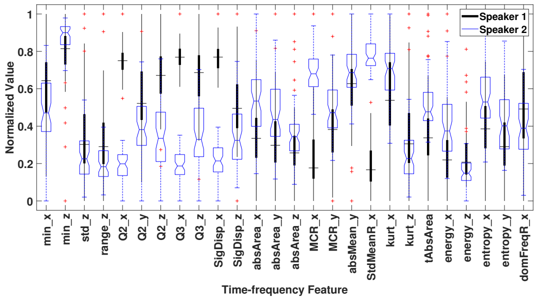


Figure 8: Illustration of the salient time-frequency features to differentiate speakers.

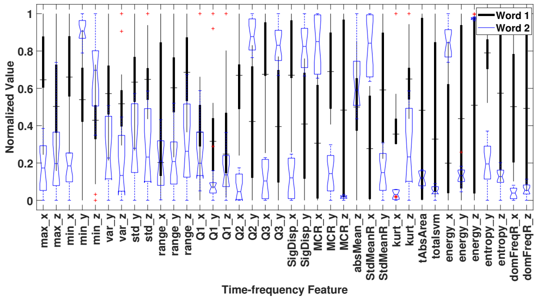


Figure 9: Illustration of the salient time-frequency features to differentiate words.

A.6 Positions of Phone's Speaker and Motion Sensor

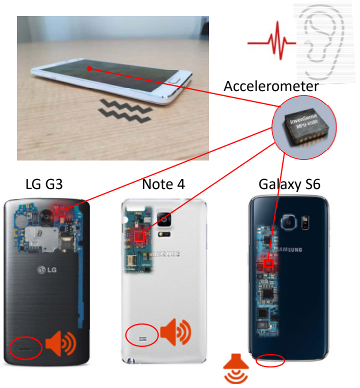


Figure 10: The speaker and the sensor positions on the smartphones of different vendors.