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

– How using the smartphone in speakerphone mode erodes your privacy –

S Abhishek Anand, Chen Wang, Jian Liu, Nitesh Saxena, Yingying Chen

***Abstract*—In this paper, we build a speech privacy attack that exploits *speech reverberations* generated from a *smartphone's inbuilt loudspeaker*[1] captured via a zero-permission motion sensor (accelerometer). We design our attack, called *Spearphone*[2], 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 successfully perform *gender classification* (accuracy over 90%) and *speaker identification* (accuracy over 80%). 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*.

I. Introduction

Today's smartphones contain a plethora of sensors aiming to provide a comprehensive and rich user experience. Some common sensors used in modern smartphones include infrared, accelerometer and gyroscope, touchscreen, GPS, camera and environmental sensors. A known security vulnerability associated with smartphone motion sensors is the unrestricted access to the motion sensor readings on most current mobile platforms (e.g., the Android OS), essentially making them *zero-permission* sensors. Recent research [1], [2], [3], [4], [5], [6] 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 [7], 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. These speech reverberations are generated due to the smartphone's body vibrating due to the principle of forced vibrations [8], behaving in a manner 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, i.e., the phone's loudspeaker, with the embedded motion sensors, it showcases the perils to a user's privacy in seemingly inconspicuous threat instances, some examples of which are 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 even the spoken words during the call (by performing speech recognition or reconstruction).
- *Speech Media Privacy Leakage:* In our attack, on-board motion sensors can also be exploited to reveal any audio/video file played on the victim's smartphone loudspeaker. In this instance, the attacker could exploit motion sensors, by logging the output of motion sensors during the media play, and learn about the contents of 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 eavesdropped media content (e.g., favorite artist).
- *Voice Assistant Response Leakage*: Our proposed threat may extend to phone's smart voice assistant (for example, Google Assistant or Samsung Bixby), that communicate with the user 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

S Abhishek Anand and Nitesh Saxena are with the University of Alabama at Birmingham. Email:$\{anandab, saxena\}$@uab.edu.

Chen Wang, Jian Liu and Yingying Chen are with WINLAB, Rutgers University. Email:$\{chenwang, jianliu\}$@winlab.rutgers.edu, $yingche$@scarletmail.rutgers.edu.

[1] Inbuilt loudspeakers are different from the earpiece speaker that is used to listen to incoming calls

[2] **Spearphone** denotes **S**peech **p**rivacy **e**xploit via **a**cclerometer-sensed **r**everberations from smart**phone** loudspeakers

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: We study the speech privacy threat that exploits the response of embedded motion sensors to the speech reverberations generated by the smartphone's loudspeakers. We carefully use existing techniques to quantify the threat by performing gender, speaker and speech classification under multiple setups. Our contributions are three-fold:

1) ***A New Speech Privacy Attack System:*** We propose a novel attack, Spearphone (Section IV), 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 which is 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 media 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* [8]. Thus, it is very hard for an attacker to eavesdrop on speech using accelerometer readings. Indeed, prior work on motion sensor exploit for compromising speech 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*. (a detailed comparison with prior work is provided in Section II). 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 III). 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. We execute the attack by carefully using off-the-shelf machine learning and signal processing techniques (Section V). 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 as discussed below and in Section V-E.

3) ***Attack Evaluation under Multiple Setups:*** We evaluate Spearphone under multiple setups mimicking near real-world usage of smartphone loudspeakers (Section VI). We show that Spearphone can perform gender and speaker classification requiring as low as just one word of test data with an f-measure $\geq$0.90 and $\geq$0.80, respectively, which shows the threat potential of the attack. 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.

Technical Challenges Addressed in Our Work: We analyzed the motion sensors' response to speech reverberations under low sampling rates and show that accelerometer is more sensitive than the gyroscope. A detailed comparison of this behavior is provided in Section III. To choose the best feature set that can accurately perform gender, speaker and speech classification, we compared the performance of the frequency-time domain features with Mel-frequency Cepstrum Coefficient features (Section V-E). For a more comprehensive speech classification and recognition, we built a word isolation and keyword search technique that could work with low sampling and fidelity sensors. We also implemented the keyword search scheme with a limited training set to approximate a harder setting for the attack (Section VI-G1).

II. Background and Prior Work

The embedded motion sensors (i.e., accelerometer and gyroscope) are useful in supporting various mobile applications that require motion tracking or motion-based command. However, they also bring potential risks of leaking user's private information. Due to the nature of the motion sensors, they can capture the vibrations associated with users' movements such as typing on the phone's keyboard. This could cause sensitive information leakage on mobile devices [9], [3], [10], [5], [11]. For instance, TouchLogger [3], TapLogger [11] and Accessory [10] utilize the accelerometer and gyroscope embedded on smartphones to infer keystroke sequence or passwords when the user inputs on the smartphone's keyboard. TapPrints [9] further shows that the tap prints on the smartphone touchscreen can be characterized by accelerometers and gyroscopes on the smartphone to identify users. In addition, (sp)iPhone [5] shows that the vibrations generated by typing on a physical keyboard can be captured by a nearby smartphone's accelerometer to derive the user's input.

Additionally, it is necessary to consider speech privacy in various daily scenarios (e.g., private meetings, phone conversations, watching or listening to media). In order to prevent unintentional listeners from overhearing the speech, traditional methods apply sound-proof walls for closed conference room to confine the speech within the room. Besides, microphone access on a smartphone is subjected to a high-level permission to prevent exploits by adversaries. In order to prevent potential snooping via smartphone's built-in microphone, people can simply deny any app's microphone permission if they are

TABLE I: Spearphone vs. prior speech privacy motion-sensor attacks/studies

	Speech Origin	Propagation Medium	Type of Vibrations	Motion Sensor	Leaked Information	Effect Present
Gyrophone	(1) Loudspeaker	Shared solid surface	Surface-aided speech vibrations	Gyroscope	Speech replayed via external loudspeakers	YES
Speechless	(1) Loudspeaker (2) Phone owner talking into his phone	(1) Shared solid surface (2) Air	(1) Surface-aided speech vibrations (2) Air-borne speech propagation	(1) Gyroscope, Accelerometer (2) Gyroscope, Accelerometer	(1) Speech replayed via external loudspeakers (2) Live human speech	(1) YES (2) Likely NOT
Spearphone	(1) Remote caller talking to phone owner, (2) Media played on phone, (3) On-board voice assistant's response	Smartphone body	Surface-borne & aerial speech reverberations	Accelerometer	(1) Remote caller's speech (2) Media played on phone (3) On-board voice assistant's response	(1) YES (2) YES (3) YES

not actively using a specific feature that requires microphone usage. The motion sensors, on the other hand, are usually freely accessible, meaning any application needs *zero permission* to access them. Additionally, MEMS sensor attributes and structures could be affected by noise and other sounds, indicating the potential to leak the smartphone user's private speech information.

Existing studies have shown that background noise affects MEMS sensors and degrades their accuracy [12], [13], [14]. The reason is that the MEMS structure can be resonant to some parts of frequencies of the sound vibrations surrounding the phone. However, due to the low sampling rate of motion sensors (e.g., 200Hz on most smartphones), its capability of snooping speech sound is often ignored or underestimated. However, the recent work shows that embedded MEMS motion sensors could reveal speech information [6], [15], [2]. Specifically, Gyrophone [6] shows that gyroscope is sensitive enough to measure acoustic signals from an external loudspeaker to reveal speaker information. Accelword [15] uses smartphone's accelerometer to extract signatures from live human voice for *hotwords* extraction.

Speechless [2] further tests the necessary conditions and setups for speech to affect motion sensors for the speech leakage. [2] concluded that motion sensors may indeed be influenced by external sound sources as long as the generated vibrations are able to propagate along the surface to the embedded motion sensors of the smartphone, placed on the same surface (*surface-aided*). [2] also showed that **aerial vibrations of speech**, such as those produced by the vocal tract of live human speakers speaking in the microphone of the phone, **do not impact its motion sensors**. Pitchin [16] presented an eavesdropping attack using embedded motion sensors in an IoT infrastructure (having higher sampling rate than a smartphone motion sensor) that is capable of speech reconstruction. They leveraged the idea of time-interleaved analog to digital conversion by using a network of motion sensors, effectively boosting the information captured by the motion sensors due to increased sampling rate obtained by sensor fusion.

The above studies, however, focus on studying the possibility or necessary conditions for making the embedded motion sensors respond to the external sound sources (e.g., loudspeaker and live human voice). Our work explores the possibility of revealing the speech played by the smartphone's built-in speakers from the phone's own motion sensors. This setting is related to a large number of practical instances, whose privacy issues are still unexplored.

Compared to the related work in [6], we found that the accelerometer performs much better than gyroscope when picking up the speech reverberations. Moreover, the study in [6] examined the speech from an external loudspeaker, which produces much stronger sound/vibration signals and only targeted the local speaker's speech using their smartphone. Smartphone loudspeakers lack the wide range of frequency response compared to an external loudspeaker (with woofers), especially at low frequencies. Since the speech signals that produce vibrations consist of low frequencies, our threat model is much weaker than the one used in [6]. Our work is not restricted to just surface-aided speech vibrations as it exploits both surface-aided and aerial vibrations that are propagated within the smartphone's own body. Thus, we believe [6] presents a threat model that is extremely favorable to the attacker but potentially too restrictive to work in the real world. A summary of related work vs. our work is provided in Table I.

In summary, in this paper, we identify and dissect live speech and media instances in which the speech privacy attack through motion sensors works, whereas a recent study [2] concluded these sensors to be "speechless" in most other setups (e.g., humans speaking into the phone, or when the loudspeaker does not share the same surface as the phone). We elaborate our detailed attack model in Section IV.

III. Motion Sensors vs. Speech Reverberations

Sound is a vibration that typically propagates as a pressure wave through a medium (e.g., air). As shown in Figure 1, when a sound is played by a mobile phone, the phone loudspeakers generate sound vibrations. Compared to the vibrations being transmitted in the air (*air-borne propagation*), the phone body also provides a pathway for propagating the resulting sound reverberations to the accelerometer and gyroscope, which are embedded in the phone (Figure 1). These embedded motion sensors are designed for sensing the physical motion of the phone, which enables various applications (e.g., fitness tracking and gaming) but they also suffer from exploitation (due to *zero-permission* nature) that draws serious security and privacy concerns.

A. Accelerometer Frequency Response

An accelerometer is an electro-mechanical device for measuring acceleration which can be either static (e.g., gravity) or dynamic (e.g., movement/vibration). The MEMS accelerometer can be modeled as a mass-spring system. An external

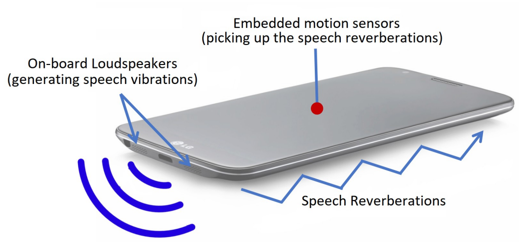

Embedded motion sensors
(picking up the speech reverberations)
On-board Loudspeakers
(generating speech vibrations)
Speech Reverberations

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

acceleration force causes movement of the tiny seismic mass inside fixed electrodes, causing a capacitive electrical signal change which can be measured as acceleration value [17].

Spearphone aims to capture the speech from the smartphone's built-in loudspeaker by leveraging the readily available motion sensors. To measure the frequency response of the accelerometer to the built-in loudspeaker sound, we play a specifically-designed signal and collect the accelerometer readings with a smartphone (e.g., Samsung Galaxy Note 4). The smartphone sensor sampling rate is set to its maximum limit of 250Hz and placed on a wood table during experiment. We generate a chirp sound signal sweeping from frequency 0Hz to 22kHz for 5 minutes and play the sound through the smartphone's built-in loudspeaker at maximum volume. This frequency range covers most of the sound range that a smartphone built-in loudspeaker is able to play. The amplitudes of the accelerometer in Appendix Figure 7(a) show that the accelerometer has a strong response to the sound frequency ranging from around 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, the spectrogram in Appendix Figure 7(b) further shows that different frequency sounds cause responses at the low frequency points of the accelerometer and generate aliased signals [6], 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.

B. Gyroscope Frequency Response

A gyroscope is a motion-sensing device used to measure device's angular velocity. The main principle of the MEMS gyroscope is the Coriolis effect, which causes an object to exert a force when it is rotating. This force can be measured by a capacitive-sensing structure supporting the vibrating mass to determine the rate of rotation. Appendix Figure 8 shows the gyroscope response to the $0-22$kHz frequency sweeping sounds from the built-in speaker. Gyroscope has observable response in the frequency range $8-9$kHz and $18-19$kHz and thus can capture some sound information in these frequency ranges. However, compared to accelerometer, gyroscope has a weaker response to the built-in loudspeaker's sound. In particular, the gyroscope shows subdued response in the frequency range $0-4$kHz (i.e., for audio sampled at 8kHz), which is more often used in practical scenarios such as telephone calls and voice messages and the speech sound lies in this frequency range. Given this property of gyroscope, we only focus on using the smartphone's accelerometer to capture the speech information.

To verify this observation, we captured a single speaker's voice in both Gyrophone [6] setup and our proposed setup as described in Section V and implemented in Section VI. The gyroscope readings' spectrum from Gyrophone setup and the accelerometer readings' spectrum from Spearphone setup are shown in Appendix Figure 9. We observed no indication of speech on Appendix Figure 9a spectrum while we noticed the speech reverberations corresponding to word "Oh" around 3.5 second mark in Appendix Figure 9b further validating our findings. We also further noted that Gyrophone setup involved a shared conducting medium that transferred the speech vibrations from the external loudspeaker to the smartphone's motion sensor. Thus, the capacity of motion sensors like gyroscope to sense these speech vibrations depends upon the nature of the shared surface. In contrast, Spearphone setup detects speech reverberations, traveling within the smartphone's body, thus is independent of any such external causes.

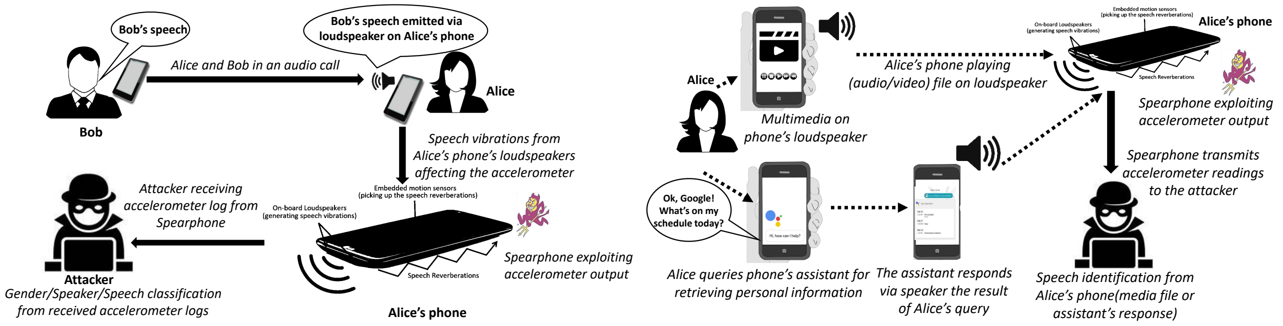
IV. Attack Overview and Threat Model

In this section, we will describe Spearphone threat model and provide an overview of Spearphone (Figure 2) that showcases the motion sensor exploiting speech reverberations. The threat model is based on [6], [2] where the embedded smartphone motion sensor readings are recorded in presence of speech in multiple setups.

A. Spearphone Threat Instances

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

- *Voice Call:* In this threat instance (Figure 2a), the victim is communicating with another person and listening in the *loudspeaker mode* (i.e not using the earpiece speaker or headphones). 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:* We also believe that the live call instance could extend to situations where human 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, inline with the victim's preferences. Malicious websites can also track the motion sensor data output in background while media content is played in the foreground. It could be a breach of privacy if a person's


Bob's speech
Bob's speech emitted via loudspeaker on Alice's phone
Alice and Bob in an audio call
Alice
Bob
Speech vibrations from Alice's phone's loudspeakers affecting the accelerometer
Embedded motion sensors (picking up the speech reverberations)
On-board Loudspeakers (generating speech vibrations)
Attacker receiving accelerometer log from Spearphone
Speech Reverberations
Spearphone exploiting accelerometer output
Attacker
Gender/Speaker/Speech classification from received accelerometer logs
Alice's phone
Alice
Multimedia on phone's loudspeaker
Alice's phone playing (audio/video) file on loudspeaker
Embedded motion sensors (picking up the speech reverberations)
On-board Loudspeakers (generating speech vibrations)
Alice's phone
Speech Reverberations
Spearphone exploiting accelerometer output
Spearphone transmits accelerometer readings to the attacker
Ok, Google! What's on my schedule today?
Hi, how can I help?
Alice queries phone's assistant for retrieving personal information
The assistant responds via speaker the result of Alice's query
Speech identification from Alice's phone(media file or assistant's response)

(a) Threat instance involving a voice call

(b) Threat instance involving multimedia/voice assistant use

Fig. 2: An overview of the proposed attack depicting possible threat instances and the attack mechanism

habits or behavior patterns are exposed to the attacker. This information could be used against the victim to discriminate them from jobs, insurance purpose, financial benefits, etc. This threat instance is depicted in Figure 2b.

- *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 (Figure 2b).

B. Attacker's Capabilities

The attacker in our threat model has similar capabilities as elaborated in previous literature [2], [6]. The attacker can fool the victim into installing a malicious application or a malicious website could track the motion sensor readings in the background via JavaScript while the unsuspecting victim is browsing. Michalevsky et al.[6] analyzed the sampling rate of motion sensors, as permissible on various browser platforms and found out that only Gecko-based browsers (e.g. Firefox) do not place any additional limit on the sampling rate of the motion sensors. Thus, the malicious attack through Javascript on a Gecko-based browser would work similar to a malicious application installed on the Android platform. These malicious applications can be designed to get triggered for specific threat instances described previously and can start logging the motion sensor output. The output can then be transmitted to the attacker where the attacker can extract confidential information.

The degree of threat posed by our attacker in Spearphone is measured by the extent of breach in speech privacy. Spearphone attempts to compromise speech privacy by performing gender, speaker, and speech classification. From an attacker's perspective, gender classification helps the attacker to narrow down the set of speakers for unidentified speech samples thereby increasing the recognition accuracy for speaker identification. Speaker classification helps the attacker with more context about the communicated speech (in addition to revealing the identity of one of the parties involved in a private voice call) while speech classification reveals the contents of the speech itself that may be considered private between the two communicating parties. 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 in scenarios where the gender of a person may be used to target them in a harmful manner. For example, advertising sites could push spam advertisements of products aimed towards a specific gender [18]. It can also be used to discriminate against a particular gender as shown in [18] where job search advertisements were gender biased. Certain oppressive societies put restrictions on particular genders and may use gender classification to target individuals in potentially harmful ways.
- *Speaker Classification (Spk-Class):* Speaker classification involves identifying a speaker that could lead to privacy leakage of the communicating parties 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 of interest 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.
- *Speech Classification (Speech-Class):* Spearphone aims to learn the actual words transmitted via the phone's speaker during the attack. In order 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 possible words from natural speech and then using word recognition on isolated words to reconstruct speech.

C. Attack Setup

The environment of the victim plays an important role in our threat model. 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 [2], [6]. 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. 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 phone is put on a table, countertop etc. in *Multimedia scenario* and *Assistant scenario*.
- *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.

Lastly, the attacker in this threat model is not in the physical vicinity of the intended victim. The attack happens through a previously installed malicious application or a rogue website that records the motion sensor data output during relevant time and sends it to the attacker. 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.

V. Attack Design

Spearphone uses a malicious application installed on the victim's phone (or through JavaScript running in a browser on the phone) to record motion sensor readings while the phone is on speaker mode. The malicious application is triggered when the victim either places a phone/video call, attempts to listen to a media file or interacts with the phone assistant.

Spearphone relies on the loudspeakers of the smartphone to generate reverberations from received speech signals. We tested the ear piece speaker, that is normally used to listen to incoming phone calls (a target for our attacker). Appendix Figure 10 shows the spectrum of the accelerometer log, recorded in the presence of an incoming voice call, that used the ear piece speaker. The call volume was set at maximum and the phone was placed on a solid surface. Appendix Figure 10 does not show any footprints of speech, indicating incapability of the ear piece speaker on LG G3 to produce speech reverberations strong enough to impact the accelerometer.

A. Motion Sensor Recording

We designed an Android application that mimics the behavior of a malicious attacker (Section IV). On start, the application immediately begins logging motion sensor readings. After a delay of five seconds from the start, we play a single word on a separate thread in the application while it 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 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 the output file containing motion sensor readings as detailed in subsequent subsections.

B. 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* as per Section IV-B. 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 III), 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 III), we try to determine the speech areas in 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 of size 100 (samples) is used. Since different words have varying lengths of utterance, we use duration of the shortest word as the size of sliding window. We calculate variance in each window to determine the behavior of the sensor within that time duration. A higher variance in the readings indicates presence of an external motion (speech vibrations). We extract the bounds of window with maximum variance as the area of sensor reading influenced due to presence of speech.

C. Feature Set for Speech Classification

Once we have extracted accelerometer readings that contain speech, we need speech features for *Gen-Class*, *Spk-Class* and *Speech-Class* that are described here. 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 IV) assumes minimal interfering noise.

Time-frequency domain features are another option to classify a signal. These features 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 converting accelerometer readings from time domain to frequency domain using Fast Fourier transformation (FFT). The FFT coefficients were used to derive energy, entropy and dominant frequency ratio that are used as frequency domain features in time-frequency features.

D. *Evaluation Metrics*

We use the following metrics to evaluate the performance of Spearphone attack: Precision, Recall, and F-measure. *Precision* indicates the proportion of correctly identified samples to all the samples identified for that particular class. In other words, 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.

E. *Design Challenges*

1) Low Sampling Rates: Operating Systems like Android place a hard limit on the data output rate for motion sensors, in order to conserve the battery life of the device. This behavior helps in freeing in valuable processing and memory power. This fact, however, makes it harder to turn the on-board motion sensors into acting as a microphone for capturing speech. Compared to an audio microphone with a sampling rate ranging from 8kHz to 44.1kHz, motion sensors become severely limited in their sampling rate (120Hz on LG G3, 250Hz on Samsung Galaxy Note 4). In addition, the on-board loudspeakers may be limited in their capacity of reproducing the audio in its true form resulting in several missing frequencies outside the loudspeaker's range. 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 III. The accelerometer response in Section III-A shows us that it was able register motion (acoustic vibrations) for the audio frequency range $100 - 3300$Hz. Comparing with gyroscope's response in Section III-B, we see that the gyroscope's response is considerably weaker than accelerometer in the human speech frequency range. Thus, we make use of accelerometer in our experiments.

2) Feature Set Selection: We compared both MFCC features and time-domain frequency features to determine the most suitable feature set that would accurately classify the speech signals captured by the accelerometer. We use the metrics as described in Section V-D and the following classifiers: Support Vector Machine (used in [6]) with Sequential Minimal Optimization (SMO), Simple Logistic, Random Forest and Random Tree (variants of decision tree classifier used in [15]). An initial experiment was conducted using the TIDigit word list [19] for using isolated words on LG G3 smartphone in *Surface* scenario. Our results indicated that time-frequency features outperformed MFCC features using 10-fold cross validation for all four classification algorithms. This result, combined with the fact that time-frequency features were proven to be efficient in [15], led us to decide upon using it in our attack for *Gen-Class*, *Spk-Class*, and *Speech-Class*. Among the classifiers, we noticed random forest outperforming other classifiers using the time-frequency features, hence we use it in the rest of our experiments (Appendix Figure 11 and Figure 12). A full set of our time-frequency features is provided in Appendix Table VIII.

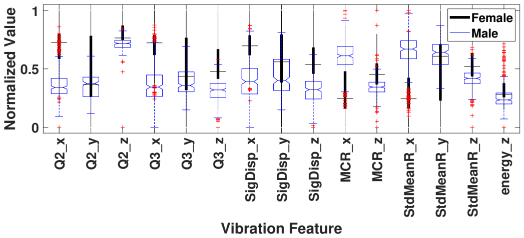


Fig. 3: Distribution of salient time-frequency features for Gen-Class.

Salient Time-frequency Features: We further studied the distribution differences of time-frequency features for *Gen-Class*, *Spk-Class*, and *Speech-Class*, because not all the features exhibit the same capability for differentiating the sound for various classification purposes. Figure 3 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 13 and 14).

3) Complete Speech Reconstruction: Performing speech reconstruction with the information captured by 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 purpose of 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 spectrogram obtained from accelerometer readings 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.

VI. Attack Evaluation

A. *Experiment Setup*

Smartphones: We conducted our experiments using three different smartphone models: LG G3, Samsung Galaxy S6 and Samsung Note 4. The experiments were performed in a quiet graduate student laboratory on a table with hardwood top for *Surface* setup, while the *Hand Held* setup was created by two participants holding the phone in their hands.

- *Operating System:* We focused mainly on phones with the Android mobile operating system as it does not require explicit user permission to obtain access to motion sensor data. In contrast, the iOS mobile operating system (from version 10.0 onwards) requires any application wishing to access motion sensor data to state its intent in the key "NSMotionUsageDescription". The text in this key would be displayed to the user describing why the application wants to access the motion sensor data. Failure to state its intent in the above described manner results in immediate closure of the said application. Also, as pointed out in Section I, 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 used in our experiments had an output data rate of 4-4000 Hz and an acceleration range of $\pm 2/\pm 4/\pm 8/\pm 16$g. The liner acceleration sensitivity range are 0.06/0.12/0.24/0.48 mg/LSB. A quick comparison with the LSM6DSL motion sensor chip used in the latest Samsung Galaxy S10 smartphone indicates similar properties for the accelerometer.

Word Datasets: *TIDigits Dataset:* We used the subset of TIDigits corpus ([19]). 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.

PGP words Dataset: We also used a pre-compiled word list 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 as detailed in Section V. We used Weka [20] as our machine learning tool 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 V-E. We used default parameters for the classification algorithm and the detailed configurations are listed in Appendix Table VI. 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., [9], [3], [10], [5], [11], [5]), including Gyrophone, have similar or even more strict training requirements for the attacker.

TABLE II: 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 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 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 and User Movement: 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. [2] and [8] claimed that external noise in human speech frequency range, traveling over the air, does not impact the accelerometer. Hence, any such noise in the surrounding environment of the smartphone would be unable to interfere with 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 be faced with in the real-world, such as during our attack instances involving phone calls.

B. Gender and Speaker Classification (Surface Setup)

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 II for *Gen-Class* and *Spk-Class*. We observe that the attack was able to perform *Gen-Class* with a substantial degree of accuracy f-measure > 0.80 with the attack being particularly successful on Note 4 as demonstrated in Table II. 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 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.

2) Surface Setup using PGP words dataset: The results with PGP words dataset for *Surface* setup, are depicted in Table II for *Gen-Class* and *Spk-Class*. Evaluating the attack against a random guess attack (0.50), we observe that the reported f-measure for the attack for all three phone models was more than 0.70 in both 10-fold cross-validation and train-test model. The attack on LG G3 boasted an f-measure of over 0.90 consistently across all the tested classification algorithms leading to the conclusion that threat measure of Spearphone when performing *Gen-Class* may indeed be harmful in this setup. Table II show Spearphone's performance when *Spk-Class* was performed using the PGP words dataset.

TABLE III: *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

For a 10-speaker classification model, a random guessing attack would give us an accuracy of 0.10. However, in our tested setup, we were able to achieve much higher f-measure scores with the attack on LG G3 achieving a score of almost 0.90. The attack on Galaxy S6 performed the worst among the attacks on all phone models but still had a better f-measure score of over 0.50 when compared to the baseline random guess attack accuracy. These results lead to conclusion that Spearphone threat is also significant while performing *Spk-Class* in this setup.

We also performed 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 speaker. 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.

C. Gender and Speaker Classification (Hand Held Setup)

1) Hand-held Setup using TIDigits dataset: Using the TIDigits dataset in *Hand Held* setup, we demonstrate the performance of the Spearphone attack in Table III for *Gen-Class* and *Spk-Class*. For *Gen-Class*, we observe that the performance of the attack on LG G3 is much 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 again significantly 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 significantly better while on Galaxy S6 it is slightly better.

2) Hand-held Setup using PGP words dataset: The *Gen-Class* attack result is shown in Table III. 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 III) shows that the attack is able to perform speaker identification with a high degree of precision 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 IV: *Effect of loudness on gender and speaker classification accuracy using Samsung Note 4 for Surfacesetup 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

D. Effect of Loudness

We further evaluate the impact of the smartphone speaker volume on the performance of Spearphone. In particular, we test the gender classification and speaker classification performance of Spearphone when setting the smartphone speaker volume to 100%, 80%, and 75% of the maximum volume. Table IV presents the results of the test on 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 very high 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 significant privacy leakage, when compared to the random guessing accuracy (i.e., 50% for gender classification and 10% for speaker classification). Moreover, people tend to use maximum volume in various scenarios to make the speech clear and comprehensible to avoid missing any important information [21]. 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 [22], [23], and speech leakage over such higher quality speakerphones could be more devastating, even at lower volume levels.

E. Result Summary and Insights

The speaker classification accuracies for Note 4 and Galaxy S6 are higher for PGP words dataset compared to TIDigits dataset. 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

TABLE V: 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

200Hz). 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).

Another interesting observation is that the *Surface* setup overall produces better classification results than the *Hand Held* setup. The minute hand motions in the *Hand Held* setup may affect the motion sensor readings and degrade the performance compared to the *Surface* setup where the phone is stationary at all times. Because the hand motions result in low frequency vibrations, we have applied a high-pass filter to reduce such influence. Another possible explanation could be the vibration absorption/dampening caused by the holding hand. To further 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%, which is 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 [24]. 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 prior work done by Michalevsky et al. [6], we find that they achieved a best case gender classification accuracy of 84% using DTW classifier on Nexus 4 which is lower than our best 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 speaker classification accuracy for Micalevsky et al. [6] was only 50% for mixed gender speakers using DTW classifier for the same dataset (TIDigits).

Natural Speech Dataset: While Spearphone achieves very high accuracy for the isolated word data set (i.e., TIDigits/PGP words), we further evaluated the performance of Spearphone with a more challenging natural speech dataset (VoxForge [25]), 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 attack is significant in a practical natural speech scenario.

F. Speech Recognition

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, which are known to be popular in the speech recognition 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.

1) Speech-Class for Single Speaker: **TIDigits dataset:** Table V shows Spearphone's accuracy of successfully 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 score of 0.74 with Random Forest classifier. In comparison, a random guess attacker would achieve an accuracy 0.09 for the tested dataset. Similar results were obtained using train-test method for classification as in Table V, though there was a slight decrease in recognition accuracy.

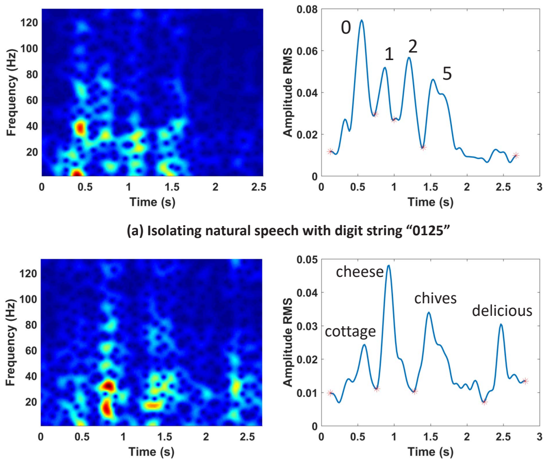
PGP words dataset: We further experimented with PGP words to explore how accurate Spearphone could recognize the isolated words other than the digits. Table V shows the *Speech-Class* results under 10-fold cross validation. By using the time-frequency features, Spearphone achieved a much higher f-measure 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.

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 further studied the feasibility of speech recognition from multiple speakers. In particular, we involve two speakers (one male; one female). Table V 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.

Gyrophone [6] 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 strongly indicate the vulnerability of smartphone's motion sensors to its own loudspeaker's speech. By combining the speech recognition and speaker identification, Spearphone is capable of further associating each recognized word to the speaker identity in multi-speaker scenarios.

G. Speech Reconstruction (Natural Speech)

We have shown the capability of Spearphone to recognize isolated words with high accuracy. To reconstruct natural



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

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

Fig. 4: Illustration of the word isolation based on the RMS of the accelerometer spectrum

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 VI-F.

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 4 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.

2) Key Word Search: Besides word isolation, 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.). The marginal words such as propositions, link verbs and other less important words can be ignored. Thus a limited-size dataset may already be sufficient for the adversary to steal most 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. Only the high confidence level predictions are kept as speech recognition results. Figure 5 shows the CDF of prediction confidence levels when $2/3$ PGP words are used as key words. We observed that we could set a threshold to differentiate the keywords from the marginal words based on the confidence level. 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 and other linguistic models for word corrections. This work is beyond the scope of our current paper and is an avenue for possible future work.

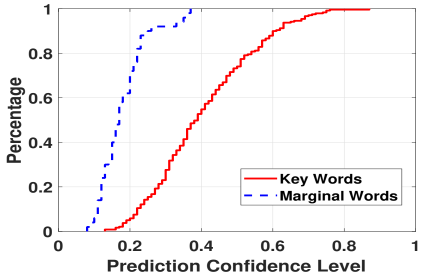


Fig. 5: CDF of the prediction confidence level for key words and marginal words

VII. Discussion and Future Work

Attack Limitations: In our experiments, we initially put the smartphone loudspeakers at maximum volume. Thus, the speech from the smartphone's loudspeakers was able to produce the strongest reverberations in the body of the smartphone, making maximum impact on the accelerometer. In reality, the loudness of different phones varies among different phone models and the loudness is also selective to each 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 attack to perform gender and speaker classification with significant accuracy, although at a lower accuracy compared to 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 downplay the perceived threat level of our attack.

Our attack targeted accelerometers embedded in the smartphones that were sensitive to the inbuilt speakers. The reverberations from the speakers travel via the body of the smartphone to the affected accelerometer. In most of the smartphones (including the smartphone models tested in this work), the motion sensor chip resides on the motherboard while the loudspeaker component is a separate unit [26]. However, all of these components are fitted in the same device tightly to reduce the overall size (thickness) of the device, leading to reverberations traveling from the loudspeaker component to the motion sensor chip.

Low sampling rate was a challenge that we faced during the implementation of our attack. Low sampling rate results in fewer data points collected by the motion sensors that directly impacts the accuracy of our attack. Resampling the obtained data to a higher sampling rate does not increase the amount of information contained in the collected data. To mitigate this challenge, we compared the accuracy of a combination of several feature sets and machine learning algorithms that maximized the amount of extracted information from our collected dataset.

Noise in the audio may be another limitation that would negatively impact the classification results of our attack. We have tried to take the noise factor into account by using Amazon Turk workers to record our speech dataset that introduces a natural level of background noise in the speech samples (that were recorded in individual Amazon Turk worker's environment). 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. However, they may not always be the case since the victim can move around with the smartphone or perform hand motions while holding the smartphone. Accelword [15] 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: Spearphone uses the smartphone's accelerometer to capture the speech of the inbuilt loudspeaker. However, the specific hardware designs of the smartphones of various vendors are different, which results in the different capabilities of the smartphone to capture the speech with accelerations. In particular, the speaker properties and the accelerometer specifications are different across various smartphone models. The specifications of the speaker and accelerometer of the three popular smartphone models are summarized in Appendix Table VII.

The accelerometers of the three models are similar but the loudspeaker of Galaxy S6 is less powerful than the others that 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 so the reverberation effect may be reduced. 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 6 shows the motion sensor specifications for some popular brands of smartphones[3]. 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.

In this work, we focused on speech reverberations from the smartphone's loudspeakers as the source of privacy leakage. While previous works exploited speech vibrations from external speech sources, Spearphone leverages the leakage of speech reverberations, that is possible due to forced vibration effect within the smartphone's body. These reverberations may be surface-aided or aerial, or a combination of both. A laser vibrometer could classify these reverberations, which will be our future work.

Accelerometer Models: The three phone models tested in this paper are embedded with the Invensense accelerometer as summarized in Appendix Table VII. We further analyzed the frequency response of another smartphone (Samsung Galaxy S3), embedding 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 types of accelerometers show the frequency range between 300Hz and 2900Hz. This indicates that the STMicroelectronics accelerometer is picking up speech reverberations similar to the tested Invensense accelerometer. 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: The design of any side channel attack exploiting motion sensors is centered around the *zero permission* nature of these sensors. To mitigate such attacks, Android platform could implement stricter access control policies that restrict the usage of these sensors. In addition, users should be made aware of the implications of permissions that they grant to applications. However, a stricter access control policy for the sensors directly affects the usability of the smartphones. Even implementing the explicit usage permission model by the applications often does not work since users do not pay proper attention to the asked permissions [27]. They often do not read all required permissions, and even when reading, they are unable to understand the security implications of granting permissions. Moreover, many apps are designed to be overprivileged by developers [28].

In addition, 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 internal build of the smartphone should be such that the motion sensors are insulated from the vibrations generated by the phone's speakers. One way to implement this approach would be to mask or dampen the vibrations leaked from the phone's speakers by surrounding the inbuilt speakers with vibration dampening material. This form of speech masking would prevent speech reverberations emanated from the phone's speakers, possibly without affecting the quality of sound

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

generated by the speaker. Speaker isolation pads are already in use in music industry in recording studios for limiting sound vibration leakage [29]. Other solutions like [30] also exist that seek to dampen the surface-aided vibration propagation that 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.

Complete Speech Reconstruction: Spearphone shows high accuracy to recognize the isolated digits/words. When applying to complete speech reconstruction, we present the initial success of isolating the words from the motion sensor readings for natural speech and the key word search to reveal sensitive information based on the limited training set. To reconstruct complete speech by combining the word isolation and key word search, further research is needed. For example, Hidden Markov Model can be used with word isolation to improve the segmentation of motion sensor readings for word. Further, linguistic models can be applied to word corrections. We leave this in our future work.

Language Identification: One possible extension of our work could be the prospect of language identification. Language identification has been performed on VoIP traffic by using the length of the encrypted VoIP packets [31]. One plausible scenario where language identification could be useful if the attacker has prior knowledge about the language preferences of the possible set of speakers to which the victim may communicate. This knowledge can help the attacker to narrow down the set of speakers. Furthermore, language can also be linked to the possible geographical location of the targeted speaker leading to privacy compromise.

Fusion of Sensors: The proposed attack exploited accelerometer that has a limited sampling rate (imposed by the operating system). It could perform better if the attacker could achieve a higher sampling rate by overriding this limit that could also be applied to gyroscope further improving the attack's performance, when combined with accelerometer's output. We leave it as future work.

VIII. 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 readings 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 (inbuilt speakers) on the phone itself. 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 transmitted through the speakers. We also discussed some possible mitigation techniques that may help prevent such attacks.

References

[1] Ahmed Al-Haiqi, Mahamod Ismail, and Rosdiadee Nordin. On the best sensor for keystrokes inference attack on android. *Procedia Technology*, 11:989 – 995, 2013. 4th International Conference on Electrical Engineering and Informatics, ICEEI 2013.

[2] S Abhishek Anand and Nitesh Saxena. Speechless: Analyzing the threat to speech privacy from smartphone motion sensors. In *Proceedings of the 39th IEEE Symposium on Security and Privacy (S&P)*, pages 116–133, 2018.

[3] Liang Cai and Hao Chen. Touchlogger: Inferring keystrokes on touch screen from smartphone motion. *HotSec*, 11:9–9, 2011.

[4] J. Mantyjarvi, M. Lindholm, E. Vildjiounaite, S. M. Makela, and H. A. Ailisto. Identifying users of portable devices from gait pattern with accelerometers. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 2, pages ii/973–ii/976 Vol. 2, March 2005.

[5] Philip Marquardt, Arunabh Verma, Henry Carter, and Patrick Traynor. (sp) iphone: Decoding vibrations from nearby keyboards using mobile phone accelerometers. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 551–562. ACM, 2011.

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

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

[8] Robert F. Coleman. Comparison of microphone and neck-mounted accelerometer monitoring of the performing voice. *Journal of Voice*, 2(3):200 – 205, 1988.

[9] E. Miluzzo, A. Varshavsky, and S. Balakrishnan. Tapprints: Your finger taps have fingerprints. In *Proceedings of ACM MobiSys*, 2012.

[10] Emmanuel Owusu, Jun Han, Sauvik Das, Adrian Perrig, and Joy Zhang. Accessory: password inference using accelerometers on smartphones. In *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications*, page 9. ACM, 2012.

[11] Zhi Xu, Kun Bai, and Sencun Zhu. Taplogger: Inferring user inputs on smartphone touchscreens using on-board motion sensors. In *Proceedings of the fifth ACM conference on Security and Privacy in Wireless and Mobile Networks*, pages 113–124. ACM, 2012.

[12] Simon Castro, Robert Dean, Grant Roth, George T Flowers, and Brian Grantham. Influence of acoustic noise on the dynamic performance of mems gyroscopes. In *ASME 2007 International Mechanical Engineering Congress and Exposition*, pages 1825–1831. American Society of Mechanical Engineers, 2007.

[13] Robert N Dean, George T Flowers, A Scotte Hodel, Grant Roth, Simon Castro, Ran Zhou, Alfonso Moreira, Anwar Ahmed, Rifki Rifki, Brian E Grantham, et al. On the degradation of mems gyroscope performance in the presence of high power acoustic noise. In *IEEE International Symposium on Industrial Electronics, ISIE 2007*, pages 1435–1440. IEEE, 2007.

[14] Robert Neal Dean, Simon Thomas Castro, George T Flowers, Grant Roth, Anwar Ahmed, Alan Scottedward Hodel, Brian Eugene Grantham, David Allen Bittle, and James P Brunsch. A characterization of the performance of a mems gyroscope in acoustically harsh environments. *IEEE Transactions on Industrial Electronics*, 58(7):2591–2596, 2011.

[15] Li Zhang, Parth H Pathak, Muchen Wu, Yixin Zhao, and Prasant Mohapatra. Accelword: Energy efficient hotword detection through accelerometer. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, pages 301–315. ACM, 2015.

[16] J. Han, A. J. Chung, and P. Tague. Pitchin: Eavesdropping via intelligible speech reconstruction using non-acoustic sensor fusion. In *2017 16th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 181–192, April 2017.

[17] Ville Kaajakari et al. Practical mems: Design of microsystems, accelerometers, gyroscopes, rf mems, optical mems, and microfluidic systems. *Las Vegas, NV: Small Gear Publishing*, 2009.

[18] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity. In *Choice, and Discrimination", Proceedings on Privacy Enhancing Technologies*, pages 92–112, 2015.

[19] Clean Digits. http://www.ee.columbia.edu/~dpwe/sounds/tidigits/, 2016.

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

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

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

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

[24] E. Khoury, B. Vesnicer, J. Franco-Pedroso, R. Violato, Z. Boulkcnafet, L. M. Mazaira Fernández, et al. The 2013 speaker recognition evaluation in mobile environment. In *2013 International Conference on Biometrics (ICB)*, pages 1–8, June 2013.

[25] Voxforge. http://www.voxforge.org/.

[26] Samsung Galaxy Note 4 Teardown. https://www.ifixit.com/Teardown/Samsung+Galaxy+Note+4+Teardown/34359, 2014.

[27] Adrienne Porter Felt, Elizabeth Ha, Serge Egelman, Ariel Haney, Erika Chin, and David Wagner. Android permissions: User attention, comprehension, and behavior. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, pages 3:1–3:14, 2012.

[28] Adrienne Porter Felt, Erika Chin, Steve Hanna, Dawn Song, and David Wagner. Android permissions demystified. In *Proceedings of the 18th ACM Conference on Computer and Communications Security*, CCS '11, pages 627–638, 2011.

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

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

[31] Charles V. Wright, Lucas Ballard, Fabian Monrose, and Gerald M. Masson. Language Identification of Encrypted VoIP Traffic: Alejandra Y Roberto or Alice and Bob? In *Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium*, SS'07, pages 4:1–4:12. USENIX Association, 2007.

Appendix

A. Classifier Configurations

TABLE VI: 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

B. Device Specifications

TABLE VII: The specifications of the speakers and motion sensors for some popular brands of smartphones

Smartphone	Motion Sensor	Output Data Rate	Phone Speaker Location
LG G3	Invensense MPU-6500	4-4000Hz	Back
Samasung Galaxy Note 4	Invensense MPU-6515	4-4000Hz	Back
Samsung Galaxy S6	Invensense MPU-6500	4-4000Hz	Bottom Edge

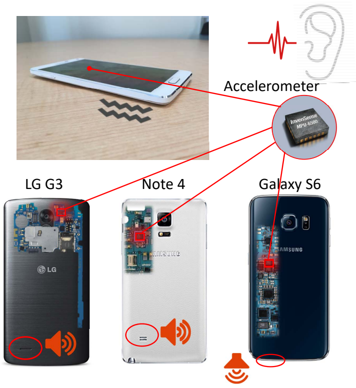
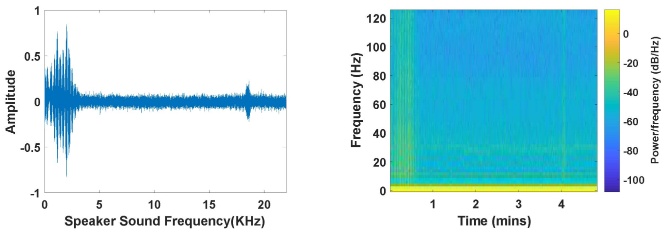


Fig. 6: The speaker and the sensor positions on the smartphones of different vendors.

C. Accelerometer Response

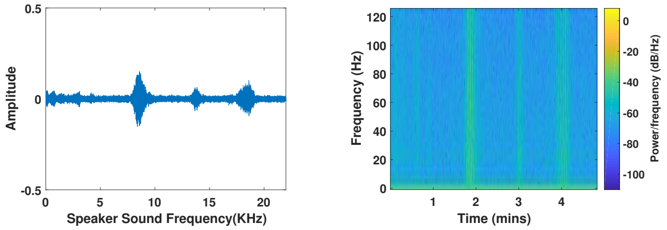


(a) Amplitude for sound 0 - 22kHz

(b) Spectrum for sound 0 - 22kHz

Fig. 7: Frequency response of the accelerometer along the z axis in response to a frequency-sweeping sound played by the smartphone's built-in loudspeaker.

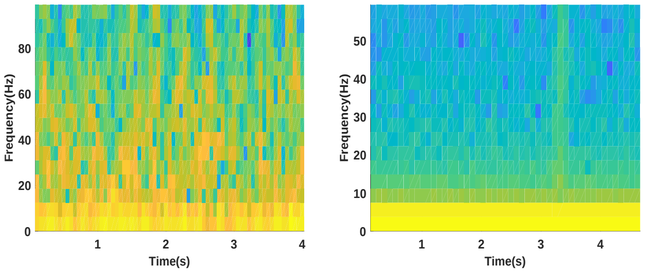
D. Gyroscope Response



(a) Gyroscope reading

(b) Spectrogram

Fig. 8: Frequency response of the gyroscope to 0 - 22kHz frequency sweeping sound



(a) Gyroscope readings' power density spectrum

(b) Accelerometer readings' power density spectrum

Fig. 9: Spectrum comparison (z axis) for the speaker "MAE" pronouncing the word "Oh" (TIDigits dataset) in [6] setup and Spearphone setup.

E. Evaluation of the Ear Piece Speaker

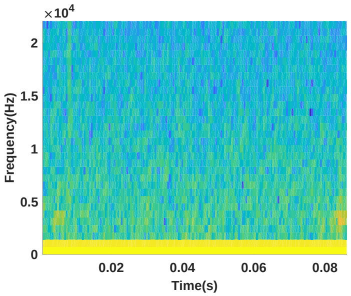
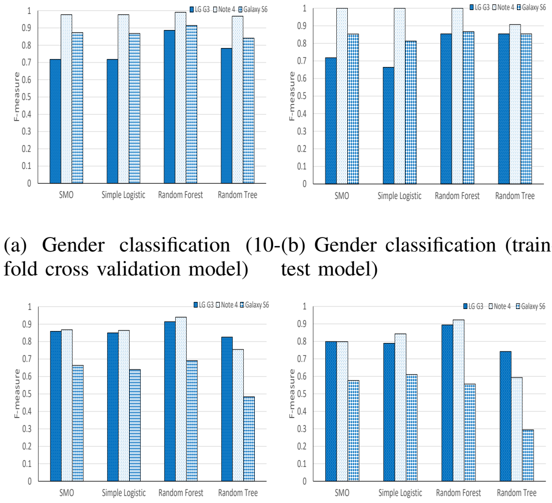


Fig. 10: Spectrum of accelerometer (LG G3) with maximum call volume on the ear piece speaker. An incoming voice call was initiated where the caller uttered digits "0" to "9" and "oh".

F. Comparison of Various Classifiers



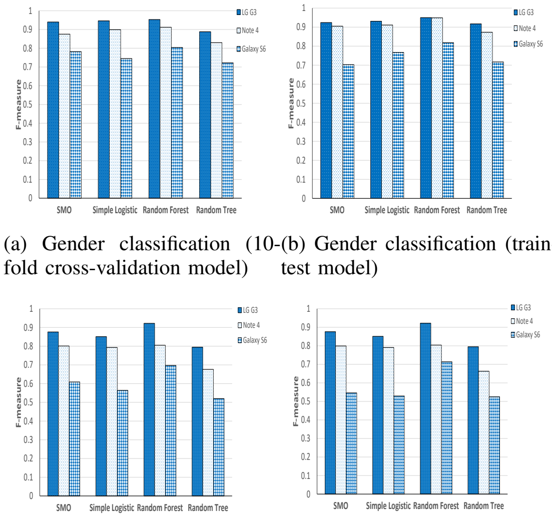
(a) Gender classification (10-fold cross validation model)

(b) Gender classification (train-test model)

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

(d) Speaker classification (train-test model)

Fig. 11: 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)

Fig. 12: Gender and speaker classification (10 speakers) for Surface setup using PGP words dataset

G. Time-frequency Feature List

TABLE VIII: 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

H. Salient Features for Speaker and Word Classification

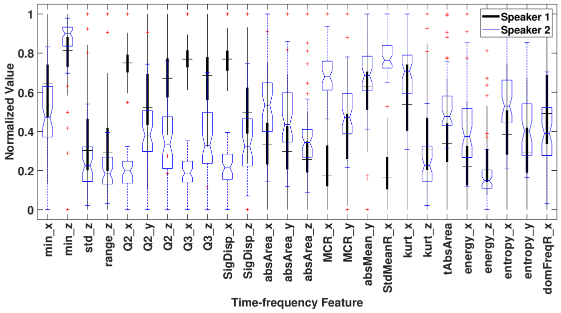


Fig. 13: Illustration of the salient time-frequency features to differentiate speakers.

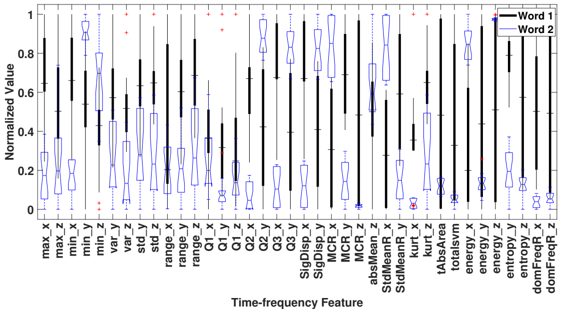


Fig. 14: Illustration of the salient time-frequency features to differentiate words.