Gyrophone: Recognizing Speech From Gyroscope Signals

Yan Michalevsky Dan Boneh
Computer Science Department
Stanford University

Gabi Nakibly
National Research & Simulation Center
Rafael Ltd.

Abstract

We show that the MEMS gyroscopes found on modern smart phones are sufficiently sensitive to measure acoustic signals in the vicinity of the phone. The resulting signals contain only very low-frequency information (<200Hz). Nevertheless we show, using signal processing and machine learning, that this information is sufficient to identify speaker information and even parse speech. Since iOS and Android require no special permissions to access the gyro, our results show that apps and active web content that cannot access the microphone can nevertheless eavesdrop on speech in the vicinity of the phone.

1 Introduction

Modern smartphones and mobile devices have many sensors that enable rich user experience. Being generally put to good use, they can sometimes unintentionally expose information the user does not want to share. While the privacy risks associated with some sensors like a microphone (eavesdropping), camera or GPS (tracking) are obvious and well understood, some of the risks remained under the radar for users and application developers. In particular, access to motion sensors such as gyroscope and accelerometer is unmitigated by mobile operating systems. Namely, every application installed on a phone and every web page browsed over it can measure and record these sensors without the user being aware of it.

Recently, a few research works pointed out unintended information leaks using motion sensors. In Ref. [27] the authors suggest a method for user identification from gait patterns obtained from a mobile device's accelerometers. The feasibility of keystroke inference from nearby keyboards using accelerometers has been shown in [28]. In [20], the authors demonstrate the possibility of keystroke inference on a mobile device using accelerometers and mention the potential of using gyroscope measurements as well, while another study [18] points to the benefits of exploiting the gyroscope.

All of the above work focused on exploitation of motion events obtained from the sensors, utilizing the expected kinetic response of accelerometers and gyroscopes. In this work we reveal a new way to extract information from gyroscope measurements. We show that gyroscopes are sufficiently sensitive to measure acoustic vibrations. This leads to the possibility of recovering speech from gyroscope readings, namely using the gyroscope as a crude microphone. We show that the sampling rate of the gyroscope is up to 200 Hz which covers some of the audible range. This raises the possibility of eavesdropping on speech in the vicinity of a phone without access to the real microphone.

As the sampling rate of the gyroscope is limited, one cannot fully reconstruct a comprehensible speech from measurements of a single gyroscope. Therefore, we resort to automatic speech recognition. We extract features from the gyroscope measurements using various signal processing methods and train machine learning algorithms for recognition. We achieve about 50% success rate for speaker identification from a set of 10 speakers. We also show that while limiting ourselves to a small vocabulary consisting solely of digit pronunciations (”one”, ”two”, ”three”, ...) we achieve speech recognition success rate of 65% for the speaker dependent case and up to 26% recognition rate for the speaker independent case. This capability allows an attacker to substantially leak information about numbers spoken over or next to a phone (i.e. credit card numbers, social security numbers and the like).

We also consider the setting of a conference room where two or more people are carrying smartphones or tablets. This setting allows an attacker to gain simultaneous measurements of speech from several gyroscopes. We show that by combining the signals from two or more phones we can increase the effective sampling rate of the acoustic signal while achieving better speech recognition rates. In our experiments we achieved 77% successful recognition rate in the speaker dependent case based on the digits vocabulary.

2 Gyroscope as a microphone

In this section we explain how MEMS gyroscopes operate and present an initial investigation of their susceptibility to acoustic signals.

2.1 How does a MEMS gyroscope work?

All MEMS gyros take advantage of a physical phenomenon called the Coriolis force. It is a fictitious force (d’Alembert force) that appears to act on an object while viewing it from a rotating reference frame (much like the centrifugal force). The Coriolis force acts in a direction perpendicular to the rotation axis of the reference frame and to the velocity of the viewed object.

Generally speaking, MEMS gyros measure their angular rate (ω) by sensing the magnitude of the Coriolis force acting on a moving proof mass within the gyro. Usually the moving proof mass constantly vibrates within the gyro. Its vibration frequency is also called the resonance frequency of the gyro. The Coriolis force is sensed by measuring its resulting vibration, which is orthogonal to the primary vibration movement. Some gyroscope designs use a single mass to measure the angular rate of different axes, while others use multiple masses. Such a general design is commonly called *vibrating structure gyroscope*.

There are two primary vendors of MEMS gyroscopes for mobile devices: STMicroelectronics [14] and InvenSense [7]. According to a recent survey [17] STMicroelectronics dominates with 80% market share. Teardown analyses show that this vendor's gyros can be found in Apple's iPhones and iPads [16, 8] and also in the latest generations of Samsung's Galaxy-line phones [5, 6]. The second vendor, InvenSense, has the remaining 20% market share [17]. InvenSense gyros can be found in Google's latest generations of Nexus-line phones and tablets [13, 12] as well as in Galaxy-line tablets [4, 3]. These two vendors' gyroscopes have different mechanical designs, but are both noticeably influenced by acoustic noise.

2.1.1 STMicroelectronics

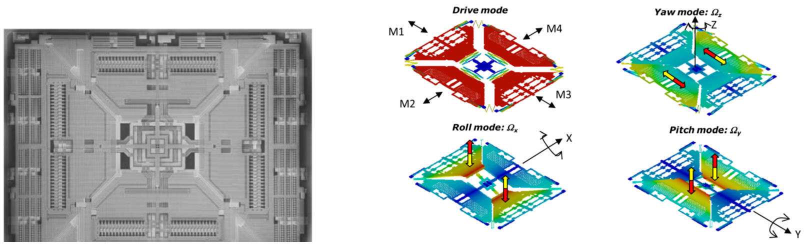
The design of STMicroelectronics 3-axis gyros is based on a single driving (vibrating) mass (shown in Figure 1). The driving mass consists of 4 parts M_1, M_2, M_3 and M_4 (Figure 1(b)). They move inward and outward simultaneously at a certain frequency[1] in the horizontal plane. As shown in Figure 1(b), when an angular rate is applied on the mass, due to the Coriolis effect, some of the masses will move (as shown by the red and yellow arrows in the figure) in a direction that is dependent on the angular rate direction.

2.1.2 InvenSense

InvenSense's gyro design is based on the three separate driving (vibrating) masses[2]; each senses angular rate at a different axis (shown in Figure 2(a)). Each mass is a coupled dual-mass that move in opposite directions. The masses that sense the X and Y axes are driven out-of-plane (see Figure 2(b)), while The Z-axis mass is driven in-plane. As in the STMicroelectronics design the movement due to the Coriolis force is measures by capacitance changes.

[1] It is indicated in [1] that STMicroelectronics uses a driving frequency of over 20 KHz.

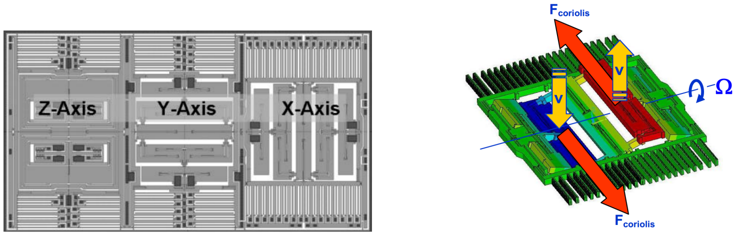
[2] According to [30] the driving frequency of the masses is between 25 KHz and 30 KHz.



(a) MEMS structure

(b) Driving mass movement depending on the angular rate

Figure 1: STMicroelectronics 3-axis gyro design (Taken from [15]. Figure copyright of STMicroelectronics. Used with permission.)



(a) MEMS structure

(b) Driving mass movement depending on the angular rate

Figure 2: InvenSense 3-axis gyro design (Taken from [30]. Figure copyright of InvenSense. Used with permission.)

2.2 Acoustic Effects

It is a well known fact in the MEMS community that MEMS gyros are susceptible to acoustic noise which degrades their accuracy [21, 22, 23]. An acoustic signal affects the gyroscope measurement by making the driving mass vibrate in the sensing axis (the axis which senses the Coriolis force). The acoustic noise has the most substantial effect when it is near the resonance frequency of the vibrating mass. Nonetheless, in our experiments we found that acoustic signals at frequencies much lower than the resonance frequency still have a measurable effect on a gyro's measurements, allowing one to reconstruct the acoustic signal.

2.3 Characteristics of a gyro as a microphone

Due to the gyro's acoustic susceptibility one can treat gyroscope readings as if they were audio samples coming from a microphone. Note that the frequency of an audible signal is higher than 20 Hz, while in common cases the frequency of change of mobile device's angular velocity is lower than 20 cycles per second. Therefore, one can high-pass-filter the gyroscope readings in order to retain only the effects of an audio signal even if the mobile device is moving about. Nonetheless, it should be noted that this filtering may result in some loss of acoustic information since some aliased frequencies may be filtered out (see Section 2.3.2). In the following we explore the gyroscope characteristics from a standpoint of an acoustic sensor, i.e. a microphone. In this section we exemplify these characteristics by experimenting with Galaxy S III which has an STMicroelectronics gyro [6].

2.3.1 Sampling

Sampling resolution is measured by the number of bits per sample. More bits allow us to sample the signal more accurately at any given time. All the latest generations of gyroscopes have a sample resolution of 16 bits [9, 11]. This is comparable to a microphone's sampling resolution used in most audio applications.

Sampling frequency is the rate at which a signal is sampled. According to the Nyquist sampling theorem a sampling frequency f enables us to reconstruct signals at frequencies of up to $f/2$. Hence, a higher sampling frequency allows us to more accurately reconstruct the audio signal. In most mobile devices and operating systems an application is able to sample the output of a microphone at up to 44.1 KHz. A telephone system (POTS) samples an audio signal at 8000 Hz. However, STMicroelectronics' gyroscope hardware supports sampling frequencies of up to 800 Hz [9], while InvenSense gyros' hardware support sampling frequency up to 8000 Hz [11]. Moreover, all mobile operating systems bound the sampling

		Sampling Freq. [Hz]
Android 4.4	application	200
	Chrome	25
	Firefox	200
	Opera	20
iOS 7	application	100 [2]
	Safari	20
	Chrome	20

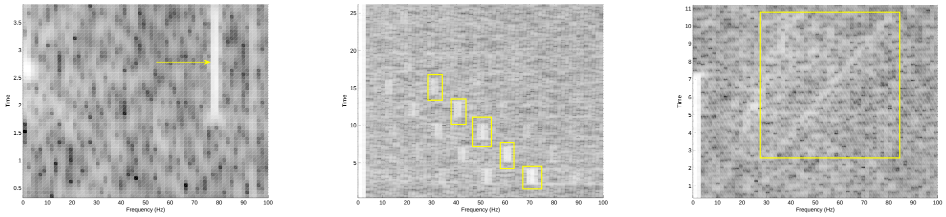
Table 1: Maximum sampling frequencies on different platforms

frequency even further – up to 200 Hz – to limit power consumption. On top of that, it appears that some browser toolkits limit the sampling frequency even further. Table 1 summarizes the results of our experiments measuring the maximum sampling frequencies allowed in the latest versions of Android and iOS both for application and for web application running on common browsers. The code we used to sample the gyro via a web page can be found in Appendix A. The results indicate that a Gecko based browser does not limit the sampling frequency beyond the limit imposed by the operating system, while WebKit and Blink based browsers does impose stricter limits on it.

2.3.2 Aliasing

As noted above, the sampling frequency of a gyro is uniform and can be at most 200 Hz. This allows us to directly sense audio signals of up to 100 Hz. Aliasing is a phenomenon where for a sinusoid of frequency f, sampled with frequency f_s, the resulting samples are indistinguishable from those of another sinusoid of frequency $|f - N \cdot f_s|$, for any integer N. The values corresponding to $N \neq 0$ are called images or aliases of frequency f. An undesirable phenomenon in general, here aliasing allows us to sense audio signals having frequencies which are higher than 100 Hz, thereby extracting more information from the gyroscope readings. This is illustrated in Figure 3.

Using the gyro, we recorded a single 280 Hz tone. Figure 3(a) depicts the recorded signal in the frequency domain (x-axis) over time (y-axis). A lighter shade in the spectrogram indicates a stronger signal at the corresponding frequency and time values. It can be clearly seen that there is a strong signal sensed at frequency 80 Hz starting around 1.5 sec. This is an alias of the 280 Hz-tone. Note that the aliased tone is indistinguishable from an actual tone at the aliased frequency. Figure 3(b) depicts a recording of multiple short tones between 130 Hz and 200 Hz. Again, a strong signal can be seen at the aliased frequencies corresponding to



(a) A single 280 Hz tone

(b) Multiple tones in the range of 130 – 170 Hz

(c) A chirp in the range of 420 – 480 Hz

Figure 3: Example of aliasing on a mobile device. Nexus 4 (a,c) and Galaxy SII (b).

130 - 170 Hz[3]. We also observe some weaker aliases that do not correspond to the base frequencies of the recorded tones, and perhaps correspond to their harmonics. Figure 3(c) depicts the recording of a chirp in the range of 420 - 480 Hz. The aliased chirp is detectable in the range of 20 - 80 Hz; however it is a rather weak signal.

2.3.3 Self noise

The self noise characteristic of a microphone indicates what is the most quiet sound, in decibels, a microphone can pick up, i.e. the sound that is just over its self noise. To measure the gyroscope's self noise we played 80 Hz tones for 10 seconds at different volumes while measuring it using a decibel meter. Each tone was recorded by the Galaxy S III gyroscope. While analyzing the gyro recordings we realized that the gyro readings have a noticeable increase in amplitude when playing tones with volume of 75 dB or higher which is comparable to the volume of a loud conversation. Moreover, a FFT plot of the gyroscope recordings gives a noticeable peak at the tone's frequency when playing tone with a volume as low as 57 dB which is below the sound level of a normal conversation. These findings indicate that a gyro can pick up audio signals which are lower than 100 HZ during most conversations made over or next to the phone. To test the self noise of the gyro for aliased tones we played 150 Hz and 250 Hz tones. The lowest level of sound the gyro picked up was 67 dB and 77 dB, respectively. These are much higher values that are comparable to a loud conversation.

[3] We do not see the aliases corresponding to 180 - 200 Hz, which might be masked by the noise at low frequencies, i.e., under 20 Hz.

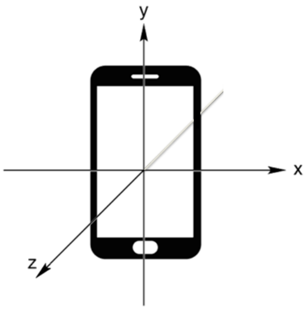


Figure 4: Coordinate system of Android and iOS.

Tone direction:	X			Y			Z		
Recording direction:	x	y	z	x	y	z	x	y	z
Amplitude:	0.002	0.012	0.0024	0.01	0.007	0.004	0.007	0.0036	0.0003

Table 2: Sensed amplitude for every direction of a tone played at different orientations relative to the phone. For each orientation the dominant sensed directions are emphasized.

2.3.4 Directionality

We now measure how the angle at which the audio signal hits the phone affects the gyro. For this experiment we played an 80 Hz tone at the same volume three times. The tone was recorded at each time by the Galaxy S III gyro while the phone rested at a different orientation allowing the signal to hit it parallel to one of its three axes (see Figure 4). The gyroscope senses in three axes, hence for each measurement the gyro actually outputs three readings – one per axis. As we show next this property benefits the gyro's ability to pick up audio signals from every direction. For each recording we calculated the FFT magnitude at 80 Hz. Table 2 summarizes the results.

It is obvious from the table that for each direction the audio hit the gyro, there is at least one axis whose readings are dominant by an order of magnitude compared to the rest. This can be explained by STMicroelectronics gyroscope design as depicted in Figure 1[4]. When the signal travels in parallel to the phone's x or y axes, the sound pressure vibrates mostly masses laid along the respective axis, i.e. M_2 and M_4 for x axis and M_1 and M_3 for the y axis; therefore, the gyro primarily senses a rotation at the y or x axes, respectively (see Section 2.1.1). When the signal travels in parallel to the phone's z axis then the sound pressure vibrates all the 4 masses

[4] This is the design of the gyro built into Galaxy S III.

up and down, hence the gyro primarily senses a rotation at both x and y axes.

These findings indicate that the gyro is an omni-directional audio sensor allowing it to pick up audio signal from every direction.

3 Speech analysis based on a single gyroscope

In this section we show that the acoustic signal measured by a single gyroscope is sufficient to extract information about the speech signal, such as speaker characteristics and identity, and even recognize the spoken words or phrases. We do so by leveraging the fact that aliasing causes information leaks from higher frequency bands into the sub-Nyquist range.

Since the fundamentals of human voices are roughly in the range of $80 - 1100$ Hz [19], we can capture a large fraction of the interesting frequencies, considering the results we observe in 2.3.2. Although we do not delve into comparing performance for different types of speakers, one might expect that given a stronger gyroscope response for low frequencies, typical adult male speech (Bass, Baritone, Tenor) could be better analyzed than typical female or child speech (Alto, Mezzo-Soprano, Soprano) [5], however our tests show that this is not necessarily the case.

The signal recording, as captured by the gyroscope, is not comprehensible to a human ear, and exhibits a mixture of low frequencies and aliases of frequencies beyond the Nyquist sampling frequency (which is 1/2 the sampling rate of the Gyroscope, i.e. 100 Hz). While the signal recorded by a single device does not resemble speech, it is possible to train a machine to transcribe the signal with significant success.

Speech recognition tasks can be classified into several types according to the setup. Speech recognition can handle fluent speech or isolated words (or phrases); operate on a closed set of words (finite dictionary) or an open set[6]; It can also be speaker dependent (in which case the recognizer is trained per speaker) or speaker independent (in which case the recognizer is expected to identify phrases pronounced by different speakers and possibly ones that were not encountered in the training set). Additionally, speech analysis may be also used to identify the speaker.

We focused on speaker identification (including gender identification of the speaker) and isolated words recognition while attempting both speaker independent and speaker dependent recognition. We did not aim to implement a state-of-the-art speech recognition algorithm, nor to thoroughly evaluate or do a comparative analysis of the classification tests. Instead, we tried to indicate the potential risk by

[5] For more information about vocal range see
`http://www.wikipedia.org/wiki/Vocal_range`

[6] For example by identifying phonemes and combining them to words.

showing significant success rates of our speech analysis algorithms compared to randomly guessing. This section describes speech analysis techniques that are common in practice, our approach, and suggestions for further improvements upon it.

3.1 Speech processing: features and algorithms

It is common for various feature extraction methods to view speech as a process that is stationary for short time windows. Therefore speech processing usually involves segmentation of the signal to short ($10 - 30$ ms) overlapping or non-overlapping windows and operation on them. This results in a time-series of features that characterize the time-dependent behavior of the signal (such as mel-frequency cepstral coefficients and short time Fourier transform). If we are interested in time-independent properties we shall use spectral features or the statistics of those time-series (such as mean, variance, skewness and kurtosis). We used MIRToolbox [24] for the feature computation.

We have tried out a few standard machine learning classifiers that have been used successfully to recognize speech. These classifiers are support vector machine (SVM), Gaussian mixture model (GMM), and dynamic time wrapping (DTW).

Prior to processing we applied to the gyroscope recordings silence removal algorithm in order to include only relevant information and minimize noise. Note that the gyroscope's zero-offset yields particularly noisy recordings even during unvoiced segments.

3.2 Experiment setup

Our setup consisted of a set of loudspeakers that included a sub-woofer and two tweeters (depicted in Figure 5). The sub-woofer was particularly important for experimenting with low-frequency tones below 200 Hz. The playback was done at volume of approximately 75 dB to obtain as high SNR as possible for our experiments. This means that for more restrictive attack scenarios (farther source, lower volume) there will be a need to handle low SNR, perhaps by filtering out the noise or applying some other preprocessing for emphasizing the speech signal.

3.2.1 Data

Due to the low sampling frequency of the gyro, a recognition of speaker-independent general speech would be an ambitious long-term task. Therefore, in this work we set out to recognize speech of a limited dictionary, the recognition of which would still leak substantial private information. For this work we chose to focus on the digits dictionary, which includes the words: zero, one, two..., nine, and "oh". Recognition of such words would enable an attacker to eavesdrop on private information, such

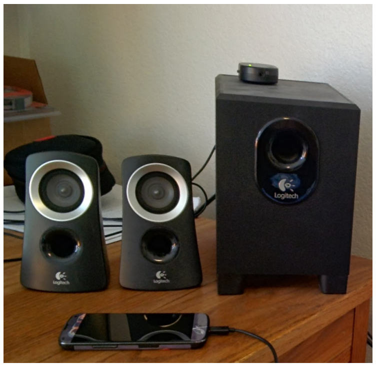


Figure 5: Experimental setup

as credit card numbers, telephone numbers, social security numbers and the like. This information may be eavesdropped when the victim speaks over or next to the phone.

In our experiments, we use the following corpus of audio signals on which we tested our recognition algorithms.

TIDIGITS This is a subset of a corpus published in [25]. It includes speech of isolated digits, i.e., 11 words per speaker where each speaker recorded each word twice. There are 10 speakers (5 female and 5 male). In total, there are $10 \times 11 \times 2 = 220$ recordings. The corpus is digitized at 20 kHz. The samples of the subset we used in our experiments can be found in ....

3.2.2 Mobile devices

We primarily conducted our experiments using the following mobile devices:

1. Nexus 4 phone which according to a teardown analysis [12] is equipped with an InvenSense MPU-6050 [11] gyroscope and accelerometer chip.

2. Nexus 7 tablet which according to a teardown analysis [13] is equipped with an InverSense MPU-6050 gyroscope and accelerometer.

3. Samsung Galaxy S III phone which according to a teardown analysis [6] is equipped with an STMicroelectronics LSM330DLC [10] gyroscope and accelerometer chip.

3.2.3 Recording utilities

To serve our experiments we implemented an Android application that samples the gyro with the highest possible frequency for a given time interval. The application code can be found at `https://bitbucket.org/ymcrcat/gyromic/src` (under the *App* directory). Additionally, to facilitate easier gyro recordings of a large number of audio files. We implemented a Python script which run on a lab computer attached to speakers and automated the process of playing each audio file in turn using the speakers, while at the same time launching the Andorid application on a USB-attached mobile device to capture the gyro recordings. The application launch was accomplished using the adb utility. The Python script, called *run_gyromic.py*, can be found under the same Git repository as the app.

3.3 Sphinx

We first try to recognize digit pronunciations using general-purpose speech recognition software. We used Sphinx-4 [31] – a well-known open-source speech recognizer and trainer developed in Carnegie Mellon University. Our aim for Sphinx is to recognize gyro-recordings of the TIDIGITS corpus. As a first step, in order to test the waters, instead of using actual gyro recordings we downsampled the recordings of the TIDITS corpus to 200 Hz; then we trained Sphinx based on the modified recordings. The aim of this experiment is to understand whether Sphinx detects any useful information from the sub-100 Hz band of human speech. Sphinx had a reasonable success rate, recognizing about 40% of pronunciations.

Encouraged by the above experiment we then recorded the TIDIGITS corpus using a gyro – both for Galaxy S III and Nexus 4. Since Sphinx accepts recording in WAV format we had to convert the raw gyro recordings. Note that at this point for each gyro recording we had 3 WAV files, one for each gyro axis. The final stage is silence removal. Then we trained Sphinx to create a model based on a training subset of the TIDIGITS, and tested it using the complement of this subset.

The recognition rates for either axes and either Nexus 4 or Galaxy S III were rather poor: 14% on average. This presents only marginal improvement over the expected success of a random guess which would be 9%.

This poor result can be explained by the fact that Sphinx's recognition algorithms are geared towards standard speech recognition tasks where most of the voice-band is present and is less suited to speech with very low sampling frequency.

3.4 Custom recognition algorithms

In this section we present the results obtained using our custom algorithm while using the three different classifiers: SVM, GMM, and DTW. We omit the from this

	SVM	GMM	DTW
Nexus 4	80%	72%	84%
Galaxy S III	82%	68%	58%

Table 3: Speaker's gender identification results

		SVM	GMM	DTW
Nexus 4	Mixed female/male	23%	21%	50%
	Female speakers	33%	32%	45%
	Male speakers	38%	26%	65%
Galaxy S III	Mixed female/male	20%	19%	17%
	Female speakers	30%	20%	29%
	Male speakers	32%	21%	25%

Table 4: Speaker identification results

white paper the technical details of the algorithms. The interested reader can find more elaborate explanations in [26]. Based on the TIDIGITS corpus we randomly performed a 10-fold cross-validation. We refer mainly to the results obtained using Nexus 4 gyroscope readings in our discussion. We also included in the tables some results obtained using a Galaxy III device, for comparison.

Results for gender identification are presented in Table 3. As we see, using DTW scoring yielded a much better success rate.

Results for speaker identification are presented in Table 4. Since the results for a mixed female-male set of speakers may be partially attributed to successful gender identification, we tested classification for speakers of the same gender. In this setup we have 5 different speakers. The improved classification rate (except for DTW for female speaker set) can be partially attributed to a smaller number of speakers.

The results for speaker-independent isolated word recognition are summarized

		SVM	GMM	DTW
Nexus 4	Mixed female/male	10%	9%	17%
	Female speakers	10%	9%	26%
	Male speakers	10%	10%	23%
Galaxy S III	Mixed female/male	7%	12%	7%
	Female speakers	10%	10%	12%
	Male speakers	10%	6%	7%

Table 5: Speaker-independent case – isolated words recognition results

SVM	GMM	DTW
15%	5%	65%

Table 6: Speaker-dependent case – isolated words recognition for a single speaker. Results obtained via ”leave-one-out” cross-validation on 44 recorded words pronounced by a single speaker. Recorded using a Nexus 4 device.

in Table 5. We had correct classification rate of $\sim$ 10% using SVM and GMM, which is almost equivalent to a random guess. Using DTW we got 23% correct classification for male speakers, 26% for female speakers and 17% for a mixed set of both female and male speakers.

For a speaker-dependent case one may expect to get better recognition results. We recorded a set of 44 digit pronunciations using a single speaker, where each digit was pronounced 4 times. We tested the performance of our classifiers using ”leave-one-out” cross-validation. The results are presented in Table 6, and as we expected exhibit an improvement compared to the speaker independent recognition[7] (except for GMM performance that is equivalent to randomly guessing).

4 Reconstruction using multiple devices

In this section we suggest that isolated word recognition can be improved if we sample the gyroscopes of multiple devices that are in close proximity, such that they exhibit a similar response to the acoustic signals around them. This can happen for instance in a conference room where two mobile devices are running malicious applications or, having a browser supporting high-rate sampling of the gyroscope, are tricked into browsing to a malicious website.

We do not refer here to the possibility of using several different gyroscope readings to effectively obtain a larger feature vector, or have the classification algorithm take into account the score obtained for all readings. While such methods to exploit the presence of more than one acoustic side-channel may prove very efficient we leave them outside the scope of this study.

Instead, we look at the possibility of obtaining an enhanced signal by using all of the samples for reconstruction, thus effectively obtaining higher sampling rate. Moreover, we hint at the more ambitious task of reconstructing a signal adequate enough to be comprehensible by a human listener, in a case where we gain access to readings from several compromised devices. While there are several practical

[7] It is the place to mention that a larger training set for speaker independent word recognition is likely to yield better results. For our tests we used relatively small training and evaluation sets.

obstacles to it, we outline the idea, and demonstrate how partial implementation of it facilitates the automatic speech recognition task.

4.1 Reconstruction algorithm

To achieve successful speech reconstruction from multiple gyroscopes we must successfully accomplish several tasks:

1. Signal offset correction – To correct a constant offset we can take the mean of the Gyro samples and compare it to 0 to get the constant offset. It is essentially a simple DC component removal.

2. Gain mismatch correction – We correct the signal gain produced by each gyro by normalizing the signal to have standard deviation equal to 1.

3. Time mismatch correction – While gyroscope motion events are provided with precise timestamps set by the hardware, which theoretically could have been used for aligning the recordings, in practice, we cannot rely on the clocks of the mobile devices to be synchronized. Even if we take the trouble of synchronizing the mobile device clock via NTP, or even better, a GPS clock, the delays introduced by the network, operating system and further clock-drift will stand in the way of having clock accuracy on the order of a millisecond[8]. One can also exhaustively search a certain range of possible offsets, choosing the one that results in a reconstruction of a sensible audio signal.

We omit the from this white paper the technical details of the algorithm. The interested reader can find more elaborate explanations in [26].

4.1.1 Evaluation

We evaluated this approach by repeating the speaker-dependent word recognition experiment on signals reconstructed from readings of two Nexus 4 devices. Table 7 summarizes the final results obtained using the sample interleaving method.

There was a consistent noticeable improvement compared to the results obtained using readings from a single device, which supports the value of utilizing multiple gyroscopes. We can expect that adding more devices to the setup would further improve the speech recognition.

[8] Each device samples with a period of 5 ms, therefore even 1 ms clock accuracy would be quite coarse.

SVM	GMM	DTW
18%	14%	77%

Table 7: Evaluation of the method of reconstruction from multiple devices. Results obtained via ”leave-one-out” cross-validation on 44 recorded words pronounced by a single speaker. Recorded using a Nexus 4 device.

5 Further Attacks

In this section we suggest directions for further exploitation of the gyroscopes:

Increasing the gyro's sampling rate. One possible attack is related to the hardware characteristics of the gyro devices. The hardware upper bound on sampling frequency is higher than that imposed by the operating system or by applications[9]. InvenSense MPU-6000/MPU-6050 gyroscopes can provide a sampling rate of up to 8000 Hz. That is the equivalent of a POTS (telephony) line. STMicroelectronics gyroscopes only allow up to 800 Hz sampling rate, which is still considerably higher than the 200 Hz allowed by the operating system (see Appendix B). If the attacker can gain a one-time privileged access to the device, she could patch an application, or a kernel driver, thus increasing this upper bound. The next steps of the attack are similar: obtaining gyroscope measurements using an application or tricking the user into leaving the browser open on some website. Obtaining such a high sampling rate would enable using the gyroscope as a microphone in the full sense of hearing the surrounding sounds.

Source separation. Based on experiments' results presented in Section 2.3.4 it is obvious that the gyro's measurements are sensitive to the relative direction from which the acoustic signal arrives. This may give rise to the possibility to detect the angle of arrival (AoA) at which the audio signal hits the phone. Using AoA detection one may be able to better separate and process multiple sources of audio, e.g. multiple speakers near the phone.

Ambient sound recognition. There are works (e.g. [29]) which aim to identify a user's context and whereabouts based on the ambient noise detected by his smart phone, e.g restaurant, street, office, and so on. Some contexts are loud enough and may have distinct fingerprint in the low frequency range to be able to detect them using a gyroscope, for example railway station, shopping mall, highway, and bus.

[9] As we have shown, the sampling rate available on certain browsers is much lower than the maximum sampling rate enabled by the OS. However, this is an application level constraint.

This may allow an attacker to leak more information on the victim user by gaining indications of the user's whereabouts.

6 Defenses

Let us discuss some ways to mitigate the potential risks. As it is often the case, a secure design would require an overall consideration of the whole system and a clear definition of the power of the attacker against whom we defend. To defend against an attacker that has only user-level access to the device (an application or a website), it might be enough to apply low-pass filtering to the raw samples provided by the gyroscope. Judging by the sampling rate available for Blink and WebKit based browsers, it is enough to pass frequencies in the range $0 - 20$ Hz. If this rate is enough for most of the applications, the filtering can be done by the driver or the OS, subverting any attempt to eavesdrop on higher frequencies that reveal information about surrounding sounds. In case a certain application requires an unusually high sampling rate, it should appear in the list of permissions requested by that application, or require an explicit authorization by the user. To defend against attackers who gain root access, this kind of filtering should be performed at the hardware level, not being subject to configuration. Of course, it imposes a restriction on the sample rate available to applications.

Another possible solution is some kind of acoustic masking. It can be applied around the sensor only, or possibly on the case of the mobile device.

7 Conclusion

We show that the acoustic signal measured by the gyroscope can reveal private information about the phone's environment such as who is speaking in the room and, to some extent, what is being said. We use signal processing and machine learning to analyze speech from very low frequency samples. With further work on low-frequency signal processing of this type it should be possible to further increase the quality of the information extracted from the gyro.

This work demonstrates an unexpected threat resulting from the unmitigated access to the gyro: applications and active web content running on the phone can eavesdrop sound signals, including speech, in the vicinity of the phone. We described several mitigation strategies. Some are backwards compatible for all but a very small number of applications and can be adopted by mobile hardware vendors to block this threat.

A Code for sampling a gyroscope via a HTML web-page

For a web page to sample a gyro the DeviceMotion class needs to be utilized. In the following we included a JavaScript snippet that illustrates this:

```
if(window.DeviceMotionEvent) {
  window.addEventListener('devicemotion', function(event) {
    var r = event.rotationRate;
    if ( r!=null ) {
      console.log('Rotation at [x,y,z] is: [' +
        r.alpha+','+r.beta+','+r.gamma+']\n');
    }
  }
}
```

Figure 6 depicts measurements of the above code running on Firefox (Android) while sampling an audio chirp 50 – 100 Hz.

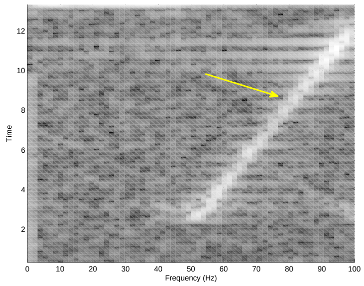


Figure 6: Recording audio at 200 Hz using JavaScript code on a web-page accessed from the Firefox browser for Android.

B Gyroscope rate limitation on Android

Here we see a code snippet from the Invensense driver for Android, taken from *hardware/invensense/65xx/libsensors_iio/MPLSensor.cpp*. The OS is enforcing a rate of 200 Hz.

```
static int hertz_request = 200;
#define DEFAULT_MPL_GYRO_RATE          (20000L)        //us
...
#define DEFAULT_HW_GYRO_RATE           (100)           //Hz
#define DEFAULT_HW_ACCEL_RATE          (20)            //ms
...
/* convert ns to hardware units */
#define HW_GYRO_RATE_NS                (1000000000LL / rate_request) // to Hz
#define HW_ACCEL_RATE_NS               (rate_request / (1000000L))   // to ms
...
/* convert Hz to hardware units */
```

```
#define HW_GYRO_RATE_HZ          (hertz_request)
#define HW_ACCEL_RATE_HZ         (1000 / hertz_request)
```

References

[1] 3-axis digital gyroscopes. http://www.st.com/st-web-ui/static/active/en/resource/sales_and_marketing/promotional_material/flyer/fl3axdigitalgyro.pdf.

[2] Corona SDK API reference. http://docs.coronalabs.com/api/library/system/setGyroscopeInterval.html.

[3] Galaxy Tab 7.7. http://www.techrepublic.com/blog/cracking-open/galaxy-tab-77-teardown-reveals-lots-of-samsungs-homegrown-hardware/588/.

[4] Inside the Latest Galaxy Note 3. http://www.chipworks.com/en/technical-competitive-analysis/resources/blog/inside-the-galaxy-note-3/.

[5] Inside the Samsung Galaxy S4. http://www.chipworks.com/en/technical-competitive-analysis/resources/blog/inside-the-samsung-galaxy-s4/.

[6] Inside the Samsung Galaxy SIII. http://www.chipworks.com/en/technical-competitive-analysis/resources/blog/inside-the-samsung-galaxy-siii/.

[7] InvenSense Inc. http://www.invensense.com/.

[8] iPad Mini Retina Display Teardown. http://www.ifixit.com/Teardown/iPad+Mini+Retina+Display+Teardown/19374.

[9] L3G4200D data sheet. http://www.st.com/st-web-ui/static/active/en/resource/technical/document/datasheet/CD00265057.pdf.

[10] LSM330DLC data sheet. http://www.st.com/st-web-ui/static/active/en/resource/technical/document/datasheet/DM00037200.pdf.

[11] MPU-6050 product specification. http://www.invensense.com/mems/gyro/documents/PS-MPU-6000A-00v3.4.pdf.

[12] Nexus 4 Teardown. http://www.ifixit.com/Teardown/Nexus+4+Teardown/11781.

[13] Nexus 7 Teardown. http://www.ifixit.com/Teardown/Nexus+7+Teardown/9623.

[14] STMicroelectronics Inc. http://www.st.com/.

[15] Everything about STMicroelectronics 3-axis digital MEMS gyroscopes. http://www.st.com/web/en/resource/technical/document/technical_article/DM00034730.pdf, July 2011.

[16] iPhone 5S MEMS Gyroscope STMicroelectronics 3x3mm - Reverse Costing Analysis. http://www.researchandmarkets.com/research/lxrnrn/iphone_5s_mems, October 2013.

[17] MEMS for Cell Phones and Tablets. http://www.i-micronews.com/upload/Rapports/Yole_MEMS_for_Mobile_June_2013_Report_Sample.pdf, July 2013.

[18] Al-Haiqi, A., Ismail, M., and Nordin, R. On the best sensor for keystrokes inference attack on android. *Procedia Technology 11* (2013), 989–995.

[19] Appelman, D. *The Science of Vocal Pedagogy: Theory and Application*. Midland book. Indiana University Press, 1967.

[20] Cai, L., and Chen, H. Touchlogger: inferring keystrokes on touch screen from smartphone motion. In *Proceedings of the 6th USENIX conference on Hot topics in security* (2011), USENIX Association, pp. 9–9.

[21] Castro, S., Dean, R., Roth, G., Flowers, G. T., and Grantham, B. Influence of acoustic noise on the dynamic performance of mems gyroscopes. In *ASME 2007 International Mechanical Engineering Congress and Exposition* (2007), pp. 1825–1831.

[22] Dean, R. N., Castro, S. T., Flowers, G. T., Roth, G., Ahmed, A., Hodel, A. S., Grantham, B. E., Bittle, D. A., and Brunsch, J. P. A characterization of the performance of a mems gyroscope in acoustically harsh environments. *Industrial Electronics, IEEE Transactions on 58*, 7 (2011), 2591–2596.

[23] Dean, R. N., Flowers, G. T., Hodel, A. S., Roth, G., Castro, S., Zhou, R., Moreira, A., Ahmed, A., Rifki, R., Grantham, B. E., et al. On the degradation of mems gyroscope performance in the presence of high power acoustic noise. In *Industrial Electronics, 2007. ISIE 2007. IEEE International Symposium on* (2007), IEEE, pp. 1435–1440.

[24] Lartillot, O., Toiviainen, P., and Eerola, T. A matlab toolbox for music information retrieval. In *Data analysis, machine learning and applications*. Springer, 2008, pp. 261–268.

[25] Leonard, R. G., and Doddington, G. TIDIGITS. `http://catalog.ldc.upenn.edu/LDC93S10`, 1993.

[26] Machlevsky, Y., Nakibly, G., and Boneh, D. Gyrophone: Recognizing speech from gyroscope signals. In *Proceedings of USENIX Security* (2014).

[27] Mantyjarvi, J., Lindholm, M., Vildjiounaite, E., Makela, S.-M., and Ailisto, H. Identifying users of portable devices from gait pattern with accelerometers. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on* (2005), vol. 2, IEEE, pp. ii–973.

[28] Marquardt, P., Verma, A., Carter, H., and Traynor, P. (sp) iphone: decoding vibrations from nearby keyboards using mobile phone accelerometers. In *Proceedings of the 18th ACM conference on Computer and communications security* (2011), ACM, pp. 551–562.

[29] Rossi, M., Feese, S., Amft, O., Braune, N., Martis, S., and Troster, G. Ambientsense: A real-time ambient sound recognition system for smartphones. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on* (2013), IEEE, pp. 230–235.

[30] Seeger, J., Lim, M., and Nasiri, S. Development of high-performance high-volume consumer MEMS gyroscopes. `http://www.invensense.com/mems/gyro/documents/whitepapers/Development-of-High-Performance-High-Volume-Consumer-MEMS-Gyroscopes.pdf`.

[31] Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., and Woelfel, J. Sphinx-4: A flexible open source framework for speech recognition. Tech. rep., 2004.