# Hello, Is It Me You're Looking For? Differentiating Between Human and Electronic Speakers for Voice Interface Security

Logan Blue
University of Florida
Gainesville, Florida
bluel@ufl.edu

Luis Vargas
University of Florida
Gainesville, Florida
lfvargas14@ufl.edu

Patrick Traynor
University of Florida
Gainesville, Florida
traynor@ufl.edu

## ABSTRACT

Voice interfaces are increasingly becoming integrated into a variety of Internet of Things (IoT) devices. Such systems can dramatically simplify interactions between users and devices with limited displays. Unfortunately, voice interfaces also create new opportunities for exploitation. Specifically, any sound-emitting device within range of the system implementing the voice interface (e.g., a smart television, an Internet-connected appliance, etc) can potentially cause these systems to perform operations against the desires of their owners (e.g., unlock doors, make unauthorized purchases, etc). We address this problem by developing a technique to recognize fundamental differences in audio created by humans and electronic speakers. We identify sub-bass over-excitation, or the presence of significant low frequency signals that are outside of the range of human voices but inherent to the design of modern speakers, as a strong differentiator between these two sources. After identifying this phenomenon, we demonstrate its use in preventing adversarial requests, replayed audio, and hidden commands with a 100%/1.72% TPR/FPR in quiet environments. In so doing, we demonstrate that commands injected via nearby audio devices can be effectively removed by voice interfaces.

## CCS CONCEPTS

• **Security and privacy** → *Access control*; Biometrics;

## KEYWORDS

Voice interface, Internet of Things

**ACM Reference Format:**
Logan Blue, Luis Vargas, and Patrick Traynor. 2018. Hello, Is It Me You're Looking For? Differentiating Between Human and Electronic Speakers for Voice Interface Security. In *WiSec '18: Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks, June 18–20, 2018, Stockholm, Sweden.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3212480.3212505

## 1 INTRODUCTION

The Internet of Things (IoT) holds the potential to increase automation in our daily lives. Devices ranging from connected appliances that report when groceries are low to smart thermostats that can anticipate desired temperature changes offer great convenience to their users. Given that many of these devices have either limited or entirely lack traditional interfaces, an increasing number now incorporate voice commands as their primary user interfaces. Voice interfaces not only simplify interaction with such devices for traditional users, but promote broader inclusion for both the elderly and those with disabilities [30].

Voice interfaces also introduce a number of security problems. First, few devices actually authenticate their users. Instead, if a command can be understood by a voice-enabled device, it simply executes the request. Any insecure sound-emitting IoT device (e.g., a networked stereo system or smart TV) near a voice interface may be used to inject commands. An adversary need not necessarily compromise nearby devices to launch a successful attack – voice controlled devices have already been intentionally [21] and unintentionally [23] activated by nearby televisions. Second, while some devices are considering the use of biometrics for authentication, this solution fails in many important cases. For instance, off the shelf tools [2, 9] allow attackers to generate audio targeting specific speakers. Moreover, even if biometrics can protect against these attacks, they do nothing to prevent against replay. Fundamentally, *wherever speakers exist, audio can easily be injected to induce voice-interfaces to perform tasks on behalf of an adversary.*

We address this problem in this paper by developing techniques that distinguish between human and electronic speakers.[1] Specifically, we identify a feature of audio that differs between the human vocal tract and the construction of modern electronic speakers. Our analysis shows that electronic speakers induce what we refer to as sub-bass over-excitation, which is the presence of very low-frequency components in the audio waveform that are not naturally produced by humans. This phenomenon is instead a consequence of the enclosures in which electronic speakers are housed. We demonstrate that this feature is a reliable indicator in detecting electronic speakers.

Our work makes the following contributions:

- **Identify sub-bass over-excitation phenomenon:** Using signal processing, we identify a frequency band present in the audio generated by electronic speakers. We discuss why sub-bass over-excitation occurs and develop the *energy balance* metric to effectively measure it.
- **Experimental evaluation of phenomenon based detector:** After explaining sub-bass over-excitation, we construct a detector that differentiates between organic and electronic speakers in low noise ($TPR$ : 100%; $FPR$ : 1.72%) and high

[1] To overcome the overloaded term "speaker", we refer to humans as "organic speakers" and manufactured devices that emit audio as "electronic speakers".


*WiSec '18, June 18–20, 2018, Stockholm, Sweden*

ACM ISBN 978-1-4503-5731-9/18/06…$15.00
https://doi.org/10.1145/3212480.3212505

noise ($TPR$ : 95.7%; $FPR$ : 5.0%) environments. We also contextualize why these false positive rates are acceptable based on reported usage data.
- **Analysis of Adversarial Commands:** We demonstrate that our detector can accurately identify the speaker as organic or electronic even in the presence of recent garbled audio injection attacks [15] and codec transcoding attacks.

We note that the sub-bass over-excitation is not simply a phenomenon limited to a certain class of electronic speakers. Rather, it is a fundamental characteristic of the construction all electronic speakers, be they of high or low quality. Without an adversary gaining physical access to a targeted environment *and* replacing the electronic speakers with custom-made devices (which, as we will explain, would add significant noise to produced audio), our proposed techniques dramatically mitigate the ability to inject commands into increasingly popular voice interfaces.

The remainder of the paper is organized as follows: Section 2 provides context by discussing related work; Section 3 offers background information necessary to understand our techniques; Section 4 details our hypothesis and assumptions; Section 5 describes the sub-bass over-excitation phenomenon and provides a confirmation test; Section 6 provides a detailed evaluation using multiple electronic speakers and test subjects; Section 7 discusses attempted mitigations and related issues and Section 8 provides concluding remarks.

## 2 RELATED WORK

Voice assistants [3, 4, 6, 7] are helpful in many ways, ranging from setting alarms or asking for the weather to making purchases or changing a thermometer's temperature. Unfortunately, these devices are vulnerable to command injection by any audio source loud enough to be heard by the built-in microphone. For example, a recent Burger King TV commercial was able to activate nearby Google Assistants [21], while a reporter triggered various Amazon Echo devices by reading a command from the teleprompter [23]. Although these incidents were not meant to cause serious harm, more severe attacks could allow an adversary to gain access to a building by opening smart locks [5], place orders [3, 8], or make financial transactions on behalf of the owner without explicit consent [10, 18]. Additionally, since voice assistants trigger with an activation phrase from any audio source, researchers have also demonstrated ways to inject malicious commands to these devices by generating sounds inaudible [37] or incomprehensible [15, 35] to humans hearing.

Determining the authenticity of the commands heard by the voice assistant is an active area of research. A commonly used method is to have a speaker recognition model that is trained using the owner's voice [19, 24]. However, having a single form of biometric authentication has been shown to have limitations due to the lack of randomness it provides for generating model [25]. Worse yet, while a trained speech model is useful, these models are also vulnerable to replay attacks [36]. Finally, generating a sound that is similar to the owner's voice has been shown capable of bypassing the speaker recognition model [22, 28]. Commercial off-the-shelf software (e.g., Lyrebird [9] and Adobe VoCo [2]) can also be used to generate fake commands akin to the voice of the owner.

To stop command replay attacks to voice assistants, liveliness verification can be performed by adding a video feed of the user's facial expression as input [11, 16, 27]. This method of proving liveliness has been widely studied in the field of information fusion [12, 17, 29]. While adding a camera source decreases the chance of malicious command injection, it also increases the chance of rejecting a real command and requires the addition of a video channel to devices, many of which do not come readily equipped with cameras. Moreover, cameras potentially introduce new threats to many environments.

In previous work, we used a secondary device controlled and colocated with the owner of a voice operated device to authenticate incoming commands [14]. However, this technique requires an additionally device which may make it unsuitable for certain applications.

In this paper, our goal is to determine whether a sound is being played through an electronic speaker or if the sound originates from organic human speech.

## 3 BACKGROUND

### 3.1 Structure of the Human Voice

Figure 1 illustrates the structures that create a human voice. The human voice is created by the complex interaction of various parts of the human anatomy. Sounds are produced by a combination of the lungs, the larynx, and the articulators (the tongue, cheeks, lips, palate, throat, and nasal cavity). The lungs force air over the rest of the vocal tract allowing it to produce sound. The larynx contains the vocal cords which are responsible for the generation of the fundamental frequency[2] present in the voice. Since the vocal cords are located at the bottom of what is essentially a closed tube, the fundamental frequency induces an acoustic resonance. This resonance generates harmonic frequencies of the fundamental frequency as it travels up and out of the human speaker's vocal tract. The articulators then alter the waveform generated by the vocal cords in order to produce the wide range of sound present in human speech. Specifically, articulators block or greatly diminish the amplitude of certain harmonics for different parts of speech. Engineers often simplify the human vocal tract into the Source-filter Model [31].

In the Source-filter Model, the human vocal tract is modeled as an underlying sound that is being filtered. Typically, women and men have fundamental frequencies between 165-255Hz and 85-180Hz respectively. [13, 34]. By generating a frequency ($x$) within a closed tube, the column of air will vibrate not just at the given frequency, but at every harmonic frequency higher than that ($2x$, $3x$, ... $nx$). The rest of the vocal tract acts as a filter, removing certain harmonics in order to produce various sounds. While this model is a very simplistic view of the biological mechanisms that produce the human voice, it will suffice for our purposes in describing the bio-mechanical system that defines the human voice.

The acoustic waveform generated by a human speaker is defined by the physical characteristics of their vocal tract. For example, men typically have larger vocal cords than women, which vibrate at a lower rate and thus cause men to have lower pitched voices.

[2] A person's fundamental frequency is the lowest frequency present in their voice.

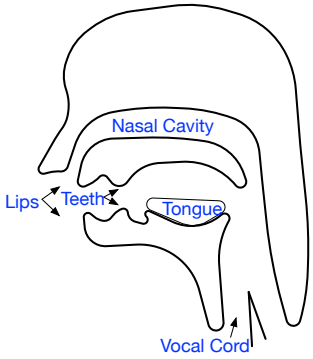



**Figure 1: Air coming from the vocal cords passes through the larynx, which generates the fundamental frequency of human voice. This sound then passes through the mouth where articulators further refine the speech. The lips, teeth, tongue, and nasal cavity make up the articulators in this figure.**

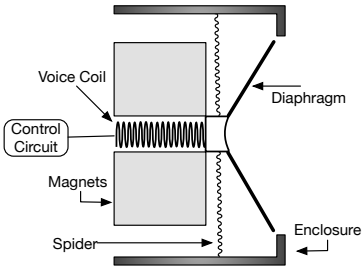



**Figure 2: To produce sound, magnets oscillate the speaker's diaphragm. The spider is responsible for connecting the speaker to the enclosure in a way that allows free movement of the speaker without separation from the enclosure.**

However, we can still make some generalization about the average human speaker despite the great variation in physical characteristics. Since articulators do not add new frequency components, the fundamental frequency is the lowest frequency that can be present in a human voice. This generalization will become the basis for our technique later in the paper.

## 3.2 Electronic Speaker Design

In Figure 2, we show a simplified diagram of a modern electronic speaker. These speakers reproduce sound by oscillating a diaphragm in an appropriate way to reproduce the recorded sound. The diaphragm displaces the air nearby and causes a pressure wave to propagate away from it. To oscillate the diaphragm, the electronic speaker uses an electromagnet, called the *voice coil*, attached to the rear of the diaphragm and located inside the magnetic field of a permanent magnet. The voice coil will induce a magnetic field when current is applied to it. The interactions between the voice coil's and the permanent magnet's fields will induce a force onto the voice coil and the diaphragm, causing it to move. The final critical component to a basic electronic speaker is the *spider*, a spring that attaches the oscillating coil and diaphragm assembly to the case. The spider must allow the diaphragm to move as freely as possible while also ensuring that the diaphragm does not separate from the case. In addition, it must ensure that the voice coil/diaphragm assembly return to its neutral point when not in the presence of a magnetic field. The material selection of the spider has a large impact on the overall performance of the electronic speaker.

An electronic speaker's design performance can be evaluated by looking at its frequency response curve. A frequency response curve describes how well an electronic speaker generates a given frequency. This curve is directly related to the physical characteristics of the electronic speaker. Namely, an electronic speaker that can accurately reproduce low frequencies will struggle to reproduce higher frequencies and vice versa. The reason that this trade off exists has to do with how energy is transferred by a wave. In order to understand why this occurs, imagine two electronic speakers, one playing a 30Hz tone and one playing a 3000Hz tone. If both electronic speakers have the same excursion (physical oscillation distance) and diaphragm size, then the electronic speaker playing the 3000Hz tone will cause the same pressure wave as the other electronic speaker 100 times more often. Since each pressure wave carries a set amount of energy, the 3000Hz electronic speaker will output 100 times more energy than the 30Hz electronic speaker and thus will be significantly louder to a listener. In order for a 30Hz electronic speaker to produce just as much acoustic energy as the 3000Hz electronic speaker, it needs to produce more energy per wave. This is possible by increasing a combination of the diaphragm's size and the excursion distance so that the 30Hz electronic speaker is displacing 100 times more air per oscillation than the 3000Hz electronic speaker. However, this has consequences on the other components of the electronic speaker. Since the diaphragm is displacing more air per oscillation, the voice coil will need to be larger to induce a stronger magnetic field and the spider will have to become stiffer to accommodate the higher amounts of momentum from the heavier voice coil and diaphragm. If the new larger electronic speaker plays a higher frequency, say 3000Hz, the higher dampening from the stronger spider would drastically reduce the amount of excursion the electronic speaker can achieve, thus reducing the amount of energy output and making higher tones significantly quieter than the lower tones. This is why many sound systems separate speakers for different frequency ranges.

Lastly, electronic speaker designers have to deal with the effects of the enclosure or case. Since every material has a resonance or natural frequency, an electronic speaker designer must account for its enclosure's vibration. Typically these enclosures resonate at somewhere in the sub-bass (20-80Hz) region. *Audio engineers design enclosures such that their resonance frequency is in this range to minimize its impact on the sound.* The sub-bass region is so low in the frequency spectrum that it is generally experienced as a pressure rather than being heard in the traditional sense. It is important to note that the enclosure will resonate whenever the electronic speaker is producing sound since it is being used as the anchor point for the spider.

# 4 HYPOTHESIS AND ASSUMPTIONS

Speech originating from an organic speaker is defined by a fundamental frequency that exists in the bass region, leaving the sub-bass

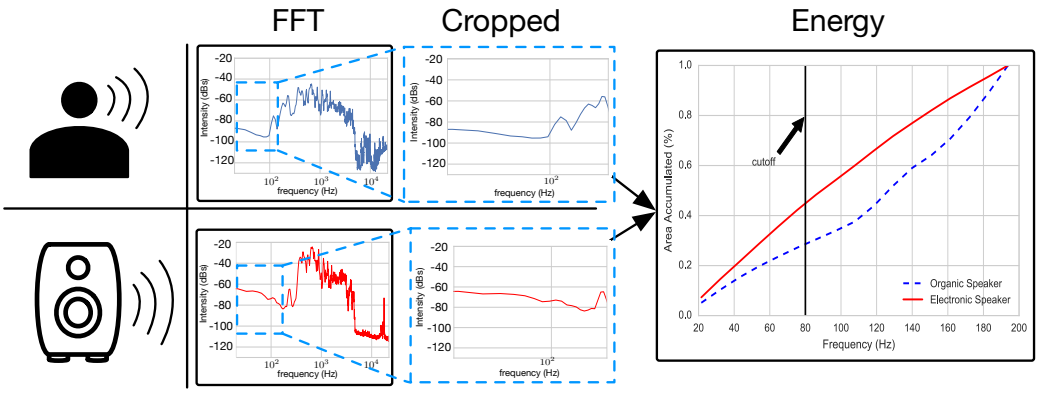



**Figure 3: This is the construction of our metric. The left most graphs show the FFT generated from the same command played via both an organic and electronic speaker. The electronic speaker has an over excited sub-bass region. The cropping in the middle graphs is done to cut down the overall area under the curves. Additionally we can see in the cropped graphs that the electronic speaker is also losing frequency components in the bass region (right side of the graph). Finally, the energy curve represents the integration and normalization of the windowed FFT. We can see that the electronic speaker accumulates energy more rapidly then the organic speaker does. Our final energy balance metric is evaluated at the cutoff frequency that divides the sub-bass and bass regions.**

region devoid of energy. In contrast, audio created by an electronic speaker will have more energy in the sub-bass region due to the resonance of the enclosure. The electronic speaker is said to have driven components in the sub-bass region since some part (the enclosure) is actively creating these frequencies. By detecting the relative amount of energy in the sub-bass region we can differentiate between electronic and organic speakers.

## 4.1 Security Model

We use the remainder of this section to describe our setting and adversary.

*Adversary:* Our adversary's goal is to inject audio commands to a target device that has a voice interface. We assume a remote adversary capable of compromising any electronic device in the room except the device implementing the voice interface. The simplest adversary would leverage devices such as TVs and radios through commercials or content that the user actively turned on. Alternatively, more advanced adversaries could comprise multiple devices around the target device. These compromised devices could then inject audio into the space without the user's consent. This type of adversary could have access to a wide range of different electronic speakers; ranging from smart TVs, computers, IP enabled webcams, or high quality speaker systems. Additionally, the adversary does not have physical access to the device. This constraint prevents the adversary from inserting their own speaker near the target device. However, we believe that this is a minor constraint since with physical access an adversary could trivially perform commands by speaking.

*Microphones:* In order to detect an electronic speaker's increased sub-bass components, our microphones must meet the following properties. First, the microphones must be capable of accurately detecting frequencies in the sub-bass region (20-80Hz). Second, the microphones must have a known frequency response curve. Microphones, just as electronic speakers, behave differently at different frequencies. By knowing the frequency response curve of the microphones we are able to compensate for any error the they may incur while recording the audio. Lastly, we require that the microphones be under our control. This requirement ensure that the data coming from the microphone has not been tampered with. Without this property an adversary could trivially defeat our technique by removing any sub-bass components before passing the audio along to be verified.

*Electronic Speakers:* In our model, the adversary can have nearly full control over the electronic speaker that is playing the audio. An adversary can control the electronic speaker's location, volume, and directionality. Additionally, an adversary could have a range of commercially available electronic speaker to be used to play the audio. As discussed in Section 3, electronic speakers are designed with audio quality in mind. This implies that all enclosures will resonate in the sub-bass region to prevent affecting any of the other more important acoustic regions. The adversary's only strict constraint is that they cannot physically alter the speaker. If an adversary altered the speaker's enclosure so that its resonant frequency moved outside of the sub-bass region, our technique could be defeated.

*Audio Sample:* We allow the adversary to have full control over the audio which is played over the electronic speaker. The adversary is free to add noise to the sample, filter out components of the sample, or change relative frequency intensities of the sample. Regardless of what the adversary does, a non-tampered speaker will still resonate in the sub-bass region more so than an organic speaker.

# 5 DIFFERENTIATING HUMANS FROM SPEAKERS

We seek to prove and measure the hypothesis stated in Section 4. The simplest way to check for sub-bass over-excitation is through visual inspection of a command's Fast Fourier Transform as can be seen in Figure 3. The Fast Fourier Transform (FFT) is an algorithm that divides a signal into its different frequency components and their amplitudes. Once again, sub-bass over-excitation is the presence of a driven component in the sub-bass region of a command. While organic speakers fundamentally lack driven sub-bass components, electronic speaker produce them due to enclosure resonance. We calculate the FFT of a command being spoken by a user and then being replayed via an electronic speaker. The sub-bass region in the command played through an electronic speaker has a higher intensity than the same region in the spoken command. In this section, we outline the construction of a metric that allows us to measure this phenomenon. Additionally, these FFTs highlight some potential complications our metric will need to overcome. However, before we examining our metric's construction we explain our experimental setup.

## 5.1 Experimental Setup

All of our commands were recorded using a far field microphone array (Respeaker 4-mic Array for Raspberry Pi) [32] that is similar to arrays in devices like the Amazon Echo [3]. For simplicity we use the Respeaker microphone array as a stand in for this devices. The Respeaker array consists of four PCB mounted microphones produced by Knowles. In comparison the Google Home and Amazon Echo Dot have two and seven PCB mounted microphones produced by TDK and Knowles respectively. Microphones can be compared via their signal to noise ration[3] (SNR). The microphones on the Respeaker array have a lower SNR (59 DBA) than both the Google Home (65 dBA) and the Amazon Echo Dot (65 dBA). From this we can discern that the microphones on the Respeaker array capture the *least* acoustic information out of the three microphones and is the *least* capable for preforming our technique.

Our microphone array recorded each of its four microphones onto a separate channel during testing. However, since our technique does not require multiple recordings, we disregard all but one of the channels. This allows for our technique to be applied to any of the aforementioned devices or any device that contains at least one microphone.

Next we will discuss our audio preprocessing that immediately follows the recording process.

## 5.2 Audio Cleaning and Preprocessing

We found that the input commands were initially noisy. Our preprocessing involved three steps: microphone equalizing, amplitude normalization, and noise filtering.

Our microphone array's microphones came with a manufacturer-provided frequency response curve. By equalizing the recorded audio with the frequency response curve of the microphone we minimized the impact they had on the recorded commands. Following the equalization, every recording was normalized (discussed in depth in Section 5.4) so that its volume was the same. This ensured that all the recordings are approximately the same intensity before processing occurred.

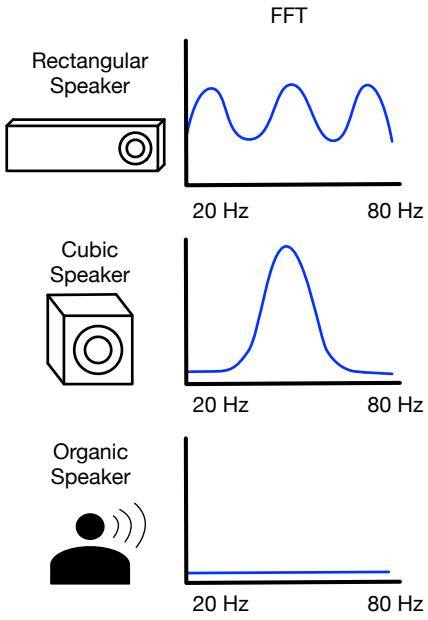



**Figure 4: The dimensions of an electronic speaker's enclosure determine the different frequencies in the sub-bass region that are over excited.**

Noise filtering was the final part of our preprocessing. We used the noise filtering function provided by Audacity [33]. The Audacity noise filter constructs a profile of the silence in the command. This profile is an averaged FFT of the silence. Then, the filter uses a sliding window over the audio to construct an FFT for each segment of audio. For each segment the filter checks if each frequency in the segment's FFT is higher than the same frequency in the silence's FFT. If so, the filter does nothing. However, if the frequency in the silence's FFT is higher, the filter subtracts a set amount (48 dBs in our case) from that frequency in the segment's FFT. This effectively brings that frequency close to zero and minimizes it's impact on the audio. It is important to note that this filter is effectively binary. For example, imagine a room had a constant 10 dBs 100 Hz noise. When a person with a voice that contains 100 Hz speaks at 40 dBs, the resulting 100 Hz component of the recording is the addition of the two sources. When the noise filter compares the silence in the room to the point when the person is speaking, it detects that the 100 Hz frequency is more intense than it was in the silence and leaves it unaltered. Both the person's and the room's components make it through the filter unchanged. Simply put, all driven frequencies in the audio will pass through the noise filter unaltered.

## 5.3 Handling Sub-bass Variation

Capturing sub-bass variation is not a straight forward process, and creates two primary challenges. The first of these complications

[3]SNR is a comparison between power of the signal and the device noise. A higher ratio means that a microphone will add less noise to the record signal; thus masking less acoustic information with noise.

is that different electronic speakers will over excite the sub-bass region differently. This reflects the design of the speaker's enclosure. Specifically, the speaker enclosure's resonant frequency is directly related to the physical dimensions of the speaker [26]. If an enclosure is rectangular, then it has three possible resonant frequencies; one for each pair of parallel walls. Each of the resonant frequencies is defined by the distance between those walls. For instance, if an enclosure was cube, then the speaker's three resonance components would all be identical. This would cause the sub-bass over-excitation to peak at a single frequency. However, if all three dimensions were different the enclosure would have three smaller resonant frequencies. This would cause the over-excitation in the sub-bass to be more evenly distributed throughout the region. This phenomenon can be seen in Figure 4. To compensate for this we designed our metric to be resilient to the variation in the sub-bass components driven by the electronic speakers.

The second complication is the variation in the sub-bass with respect to time. That is, an electronic speaker may produce more sub-bass during a given phoneme of a command than another. This is due to how the different frequency component of a phoneme excite the enclosure's resonance. Simply put, certain frequencies will cause the enclosure to resonant more than others. A command recorded from an organic speakers may also contain additional sub-bass from various backgrounds sources. Sources including bumps and knocks in the background can cause higher than normal amounts of sub-bass to be present. These temporary inconsistencies in sub-bass will cause the command's FFT to misrepresent the amount of sub-bass present in the commands. Once again, our metric was constructed in such manner so that it is robust to this complication.

## 5.4 Metric Construction

**Step 1:** The construction of our metric begins with applying a sliding window to the recorded audio. The window size was selected to be 0.1 seconds long with no overlap. We address window size selection shortly. By applying a sliding window to the input audio, we make our metric robust against sub-bass variation with respect to time. We do this by computing a metric for every window and normalization of the sample at the end. This normalization step will be discussed at the end of this section. Figure 3 shows an overview of how our metric is calculated for a single window.

**Step 2:** Next we compute an FFT with 4096 output points for each window. FFTs average frequency amplitude across the sample they are applied to which makes them susceptible to outliers. By windowing the audio command we can prevent outlying amounts of sub-bass from background noises or certain phonemes from skewing our data. Once again, this is handled by the normalization at the end of this section. Each FFT is then cropped down to contain only frequencies between 20 and 250 Hz. There is a trade off between the size of the sliding window and the FFT. The larger the FFT the more data points we have within our cropping frequency range. However, larger FFTs require more audio samples (time) as input and become more susceptible to outliers in the sub-bass region. Our window and FFT size selection allows us to maintain a large enough number of points in this frequency range (21 points) while having a short enough time window to become robust to sub-bass outliers.[4] The cropping of the FFT makes changes in the sub-bass region easier to detect. This statement will be clarified later in this section.

**Step 3:** Next, we integrate over the cropped FFT to create a spectral energy curve. The spectral energy curves represents the total energy of the audio in the 20-250 Hz range. This curve is then normalized so that the area underneath the curve is equal to one. This makes the value at any point along the curve equal to the cumulative distribution function.

**Step 4:** We define a cutoff value where the normalized energy curve is evaluated. In other words, we selected a point along the curve that defines the separation of the sub-bass and bass regions. An example cutoff value can be seen in the last panel of Figure 3. In our tests we chose a cutoff value of 80 Hz. At that point, our normalized energy curve evaluates to the total percentage of energy that is present in the sub-bass. This is equivalent to the following equation:

$$energy\ balance\ metric = \frac{E_{Sub\text{-}bass\ Region}}{E_{Total\ Evaluated\ Region}} \quad (1)$$

where $E_{Sub\text{-}bass\ Region}$ represents the energy accumulated in the sub-bass region and $E_{Total\ Evaluated\ Region}$ is the energy accumulated in the whole region being examined (20-250 Hz). Examining the sub-bass region in this way, allows our metric to be robust against the various different enclosure shapes as described in Section 5.3. *Whether the sub-bass over-excitation is spread out or concentrated into a single peak, the amount of energy present in that region will remain approximately the same.*

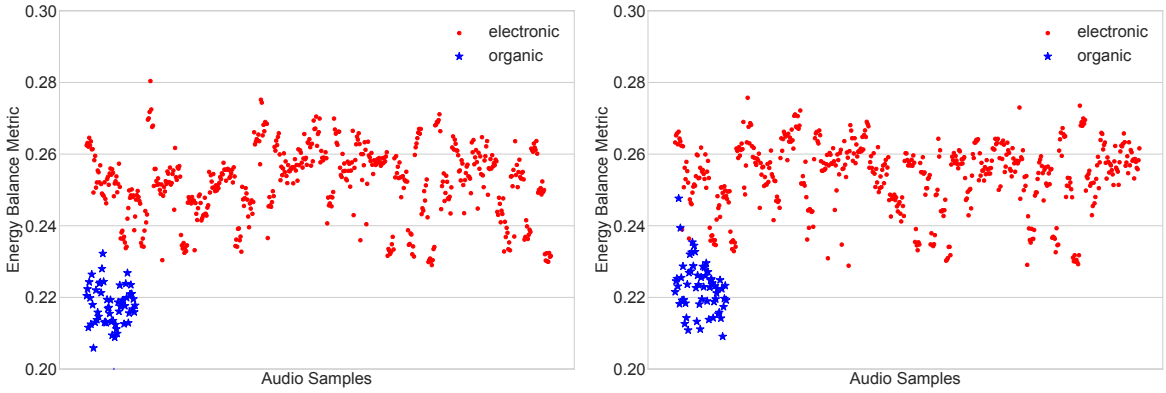
It is at this point that the earlier cropping of the FFT has an impact. By cropping the FFT, the sub-bass region becomes a larger portion of the total energy represented. This means that smaller changes in the amount of spectral energy in the sub-bass region will result in larger changes to the normalize energy balance. Additionally, the FFT cropping allows us to pick up on a second phenomenon that is common with small and low end speakers: they struggle to reproduce bass frequencies. This means that electronic speakers produce too much energy in the sub-bass region, while simultaneously having too little energy in the bass region. This causes the energy curve from an electronic speaker to further deviate from that of an organic speaker.

**Step 5:** Finally, our metric must handle variation in the sub-bass with respect to time as discussed earlier in this section. In order to prevent outlying sub-bass components from affecting the final energy balance, we fit the accumulated energy balances to an normal distribution by removing outliers based on the skewness of our data. Once our data's skew is approximately zero, we select the median value from the data as our final energy balance metric.

# 6 EVALUATION

We evaluate the performance of our normalized energy balance metric. For testing we collected samples from eight human speakers, four male and four female. We believe that eight speakers is sufficient given that our technique is not attempting to identify the individual speakers. We include both male and female speakers to ensure we have a wide range of speaker pitches. To properly validate our energy balance metric we instead need a large amount

[4] The average phoneme length for a speaker falls somewhere between 100-300ms [20].




(a) Low Noise Environment (b) High Noise Environment

**Figure 5: The energy balance metric for both environments shows similar values. However, organic speakers show a higher energy value in high noise environments than in low noise environments. The increase in energy is an artifact of the noise filtering process having to remove large amounts of background noise.**

of spoken phrases. To satisfy this, we recorded each speakers speaking eight different command phrases.[5] These commands were then recorded and played back through eight commercially available electronic speakers that capture a range of different speaker :x

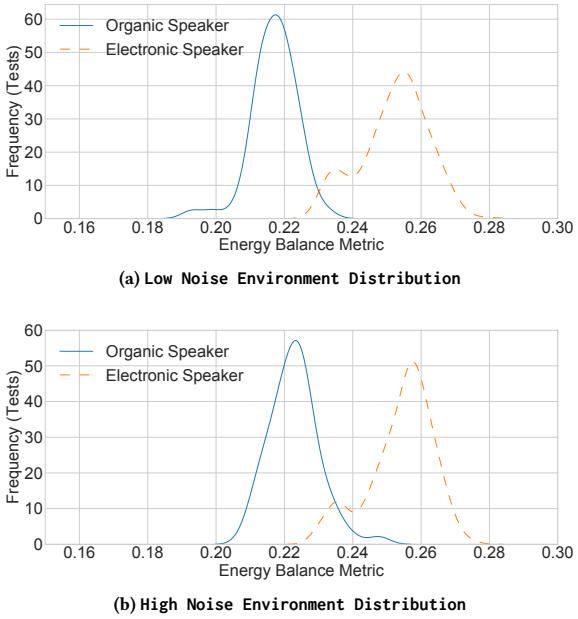
speakers used including (A) Music Angel, (B) HP Computer Speakers, (C) Dynex TV, (D) Acer Predator monitor speakers, (E) Samsung Sound Bar, (F) Insigna TV, (G) MacBook Air, and (H) Bose Wave. To determine the effects of background noise on our detection method, we repeated each of the commands in environments that had low (e.g., normal bedroom) and high (e.g., office space with white noise generators at 50dB and scattered conversations) background noise levels. In total we obtained 1204 samples: 605 sample points (58 for organic and 547 for electronic speakers) in a low background noise environment and 599 (60 for organic and 539 for electronic speakers) in a high background noise environments.[6]

We contacted our Institutional Review Board (IRB) regarding the use of human voices. Because the analysis in this work was effectively about electronic speakers and not the people themselves, they indicated that *no further IRB review or approval was necessary.*

## 6.1 Calculation Threshold and Performance Evaluation

Figure 5 shows the energy balance metric (derived in Section 5.4) for each sample in both testing environments. A qualitative analysis of these graphs shows that organic speakers are more likely to have a lower energy balance than electronic speakers. To determine if the audio sample comes from an organic or electronic speaker, a detector can be built around this phenomena by choosing an optimal threshold limit as a minimum value for electronic speakers.

Before evaluation our detector, we need to derive a threshold limit for our energy balance metric to determine if the audio is coming from an organic rather than an electronic speaker. Figure 6 shows the distribution of the energy balance metric that comes




(a) Low Noise Environment Distribution

(b) High Noise Environment Distribution

**Figure 6: The distributions of the calculated energy balance metric between organic and electronic speaker are different in both environments tested.**

from both organic speakers and electronic speakers in both testing environments. Since there is an overlap in the distributions for both environments, determining an optimal threshold for our metric requires a trade off between false positives (i.e., organic speakers identified as electronic speakers) and true positives (i.e., electronic speakers identified as electronic). To do that, we calculated ROC curves, which give us the performance of our detector under various threshold limits. Figure 7 shows the performance trade off of the detector in environments with low and high background noise levels. The accuracy of distinguishing between organic and

[5]We placed our command phrases in Appendix A.
[6]We discarded audio samples that had errors in recording which were discovered after data collection.

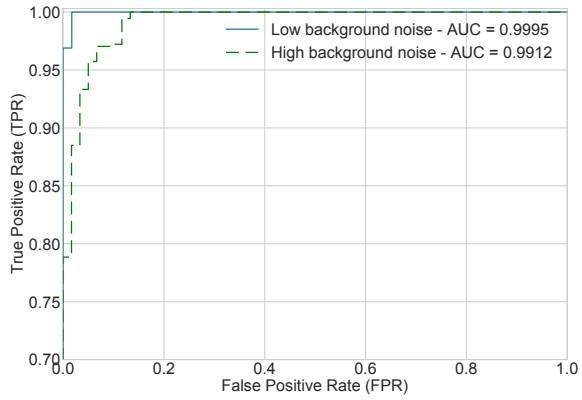



**Figure 7: The ROC curves for the speaker detector shows that our metric can easily differentiate between organic and electronic speakers in environments with low and high levels of background noise.**

electronic speakers can be measured by calculating the area under the curve (AUC) of each ROC curve, with AUC values closer to 1.0 signifying better performance. From this figure, we calculated the AUC of both environments to be 0.9995 and 0.9912 for low and high noise levels respectively. These values tell us that our detector has excellent accuracy in terms of distinguishing between organic and electronic speakers at a wide range of threshold values. However, since our use case application is to prevent injection attacks to voice assistants, we must optimize to have a high true positive rate (TPR) while still retaining reliable performance (i.e., not stopping actual organic commands). We define a reliable performance as having a false positive rate (FPR) no greater than 5.0%, which equates to 1 every 20 organic commands being rejected. For reference, most voice assistant users place four voice commands on a daily basis [1]. With the currently set FPR, these users can expect command rejection *once every five days*. We believe this is a reasonable trade-off because when a command is rejected, the user can simply repeat it.

In Figure 8, we show the performance output for possible energy balance threshold limits. For low noise environments, we choose a threshold value of 0.2280 and achieve a FPR of 1.72% while having a TPR of 100.0% (Figure 8a). For reference, by choosing this threshold for our energy balance metric, we would correctly stop all injection attacks coming from electronic speakers while minimizing the performance degradation of voice assistants by only stopping 1 every 58 organic voice commands (once every two weeks).

For high noise environments, we choose an energy balance threshold limit of 0.2345, and achieve our performance reliability FPR of 5.0%. However, in this environment our TPR decreases to 95.7%. The drop in accuracy can be attributed to a decrease of performance in organic speakers rather than an increase of performance in electronic speakers. We believe the increase in FPR is due to the noise filter used in preprocessing, which removes bass components in the organic speakers voice. As we mentioned in Section 5.2, noise filtering is a crucial step of our detection mechanism and is binary by nature: if a component of an organic speaker was unable to break the intensity threshold, it was removed. Since female speakers generally contain less intense bass components, we believe the filter removed all traces of the bass components from

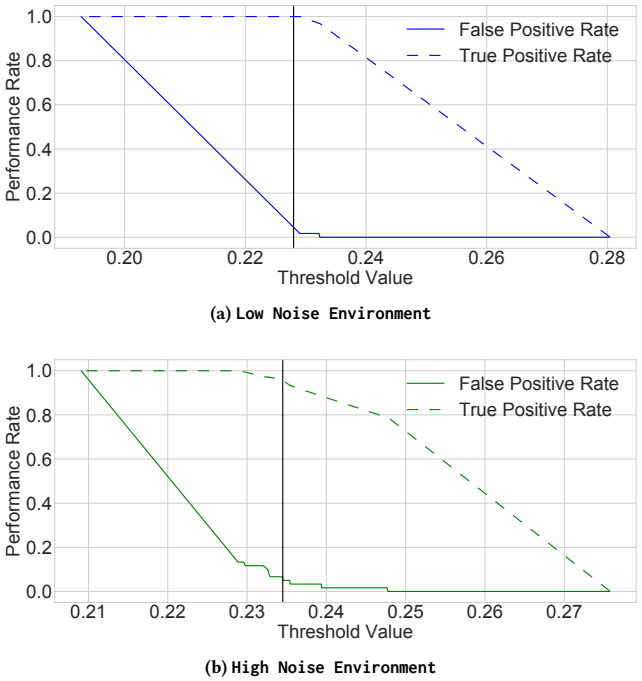



**Figure 8: Picking a threshold level requires a trade off between performance and detection rates. These graphs show the True and False Positive Rates for both environments over a range threshold values. We chose our thresholds (the vertical black line) to reach a performance rate of 95% while keeping a false positive rate to a minimum.**

their voices, which caused their energy balance metrics to skew higher. If true, then we would expect the male speakers to have a similar performance as before. To test this hypothesis, we used the highest male energy balance as a threshold and reanalyzed the data to get a TPR of 99.2%. This TPR confirms the drop in accuracy was caused by the preprocessing filter in a high noise environment. This accuracy can be maintained by simply having the speakers speak loud enough to overcome the effects of the noise (i.e., setting a minimum volume threshold for acceptable commands).

## 6.2 Statistical Significance

We showed how our technique is able to differentiate audio between electronic speakers and organic speakers. To demonstrate that our differentiation technique is statistically significant, we use an independent sample $t$-test. This is a null hypothesis test that determines if the mean of two populations are equal. For our experiments, the null hypothesis would indicate that our technique does not differentiate between organic and electronic speakers in a statistically significant way.

We separated our data by environments (low and high background noise), set our confidence interval to $\alpha = 0.05$, and then performed the test on each environment. We also calculated r-effect, which tells us the strength of our findings (with $> 0.5$ meaning large effect) and Cohen-d, which tells us how large the effect of the phenomena would be (with $> 0.8$ meaning large).

*Low Background Noise.* In total we had 58 organic and 547 electronic samples. The Cohen-d value for our sample sets was 4.16 with an r-effect of 0.901 indicating a large effect size and our calculated p-value was $< 0.001$ with a power of 1. These results demonstrate an extreme likelihood of a statistically significant difference between both sample sets. Since the populations' means differ, we can reject the null hypothesis and confirm our results are statistically significant.

*High Background Noise.* In total we had 60 organic and 539 electronic samples. The Cohen-d value for our sample sets was 3.71 with an r-effect of 0.880 indicating a large effect size and our calculated p-value was $< 0.001$ with a power of 1. These results demonstrate an extreme likelihood of a statistically significant difference between both sample sets. Since the populations' means differ, we can again reject the null hypothesis and confirm our results are statistically significant.

## 6.3 Adversarial Input

We tested our detector with two different attack vectors for voice command injections.

*Hidden Commands.* We passed audio samples from 10 different hidden commands [15] that were provided to us by the authors. These audio files were specifically made to trick voice assistants to recognizing commands even if the commands themselves were not discernible to humans. Since the audio samples were artificially created, we could only play them through electronic speakers (rather than organic). We again tested samples[7] in environments with low and high background noise levels. The minimum value for the energy balance metric for the adversarial samples was 0.2601 (shown in Figure 9). By using the threshold limits derived earlier in this section, we were able to correctly detect (and reject) each audio sample as an electronic speaker.

*Codec Transcoding Attacks.* We used lossless `wav` encoding in all previous experiments. However, an adversary may attempt to bypass our mechanism by using alternate encoding schemes. For instance, an adversary can inject commands to voice assistants by playing the command itself from a phone call. In this case, because the audio comes from a phone call, the audio must first go through a codec that compresses the audio before it gets sent through the cellular network. Alternatively, because GSM codecs remove both high and low frequencies, an adversary may believe this approach can fool our detector. To test our detector under this attack vector, we passed a sample set of our collected audio through a GSM-FR codec and then measured the energy balance metric of the compressed audio. Again, in Figure 9, we show the energy balance value for each compressed audio sample. These samples are easily detected even with energy balance limit set to the derived high noise threshold.

## 6.4 Evaluation Summary

In Section 4, we set out to determine whether audio originates from an organic speaker or an electronic speaker by measuring the energy balance of an audio sample. Our evaluation shows that after deriving the energy balance threshold for low noise environments, our detector is able to correctly detect audio with a false positive rate of 1.72%. Similarly, after deriving the energy threshold for high noise environments, our detector correctly detects audio with a reasonable false positive rate of 5.0%. *These values correspond to days or weeks between falsely rejected commands.* Finally, the best available adversarial evasion techniques were also easily defeated.

[7]For this analysis, we played the sound through the (A) Music Angel and (B) HP computer speakers.

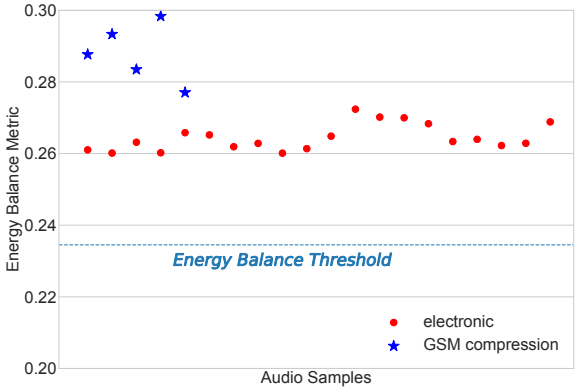



**Figure 9: The energy balance metric is able to correctly detect all adversarial input as electronic speakers.**

# 7 DISCUSSION

## 7.1 Limitations

All experiments were performed with commercial off-the-shelf speakers designed for optimal sound quality. An electronic speaker that was not designed for sound quality could produce resonance frequencies that would cause a false negative; however, the goal of commercial electronic speakers is to reproduce sound for human hearing. In other words, a speaker that could reliably be misclassified by our metric would be intentionally manufactured to do so. This specially designed speaker would produce low-quality/high noise audio making it unlikely to be purchased by customers or sold by a reputable manufacturer. An adversary would therefore have to manufacture the speaker themselves. While homemade speakers are common today, and adversary would still need to place their speaker near the device they are trying to compromise. In order to place their speaker, an adversary would need physical access to the target device which is outside our adversarial model as defined in Section 4.1.

An adversary could also potentially modify an existing speaker to reliably misclassify, however this would have adverse effects on the sound quality and be noticeable to a user. Additionally, altering the speaker would not be feasible for many adversaries [21, 23] given the complex nature of speaker design. Altering a speaker that is already located near the target device would also require physical access to the location of the speaker. Once again, the need for physical access is outside the capability of our adversary.

Our technique is also vulnerable if an adversary can control the silence that our noise filtering is based on. An adversary could construct the silence to contain large amounts of energy in the sub-bass region. This would cause our noise filter to designate higher amounts of sub-bass as noise, potentially masking and then

removing an electronic speaker's over-excitation. To prevent such an attack, we could sample the silence during device's initial setup. This constricts the attack window of an adversary to the short initialization phase of the device's life. If the background noise around the device later changes, the user would have to reinitialize the silence sampling in a secure manner. In practice, this could be done via a physical button or smartphone application.

### 7.2 Speaker Quality

Our experiments were performed using a wide range of different speakers. We believe that the speakers used are representative of a wide range of commercially available speaker. Broadly, electronic speakers can be broken into two distinct categories, single driver and multi-driver systems. Single driver systems contain electronic speakers that are responsible for producing the full frequency range. In contrast, multi-driver systems have different electronic speaker dedicated for different frequency ranges. Our test set included both classes.

*Single Driver Speaker Systems.* Single driver systems are common in devices that are inexpensive or more size constrained. We expect the vast majority of speakers in IoT devices and smartphones to fall in this category. In our testing, the Music Angel (A), Dynex TV (C), Acer Predator Computer Monitor (D), Insignia TV (F), and MacBook Air (G) (Appendix B) are all single driver systems. As discussed in Section 3, different frequency ranges require different physical characters to produce. As a result, single driver systems have frequency response curves with more variance and struggle to produce intense bass components. In addition to the electronic speaker's sub-bass over-excitation, our energy metric also captures the lack of bass in the audio. The decreased amount of bass components will make the sub-bass contributions appear larger, thus increasing the detectability of the speaker. Due to their lack of bass and sub-bass over-excitation, single driver speakers are the easiest for our metric to detect. Additionally, these types of speakers are the most likely to be compromised by an adversary given their extensive use in commodity devices.

*Multi-driver Speaker Systems.* Multi-driver systems are common in more expensive and dedicated electronic speakers. These systems contain dedicated speakers designed to produce different frequencies ranges, the most common of which is a bass speaker to produce low frequencies. The HP Computer Speakers (B), Samsung Sound Bar (E), and Bose Wave IV (H) (Appendix B) from our testing are all multi-driver systems. Dedicated bass speaker enclosures can be broken into either ported (HP Computer Speakers and Samsung Sound Bar) or non-ported (Bose Wave IV) designs. Ported speakers[8] are the more common of the two types, with non-ported speakers generally only being used in single enclosure devices like the Bose Wave IV. Ported bass speakers are designed to *increase* the amount of sub-bass over-excitation generated by the speaker. The port amplifies the case's resonance frequency to create a more "powerful" bass notes that can be felt by the listener. As a direct result of this the sub-bass region is over-excited more for a ported bass speaker than a non-ported bass speaker.

[8]Ported speakers have a large open hole or "port." These are most often found on subwoofers.

Additionally, multi-speaker systems usually have flatter, more natural frequency response curves. Their improved frequency response characteristics could make them harder for our technique to detect. However, ported bass speakers are common amongst high end multi-driver speaker systems. As a result, our technique can easily detect these kinds of systems due to the intentional amplification of the sub-bass region.

In contrast, non-ported bass speakers do not amplify their sub-bass region intentionally. This makes non-ported dedicated bass speakers the hardest for our technique to detect. In order to detect a non-ported bass speaker we must identify only the non-amplified sub-bass over-excitation. In our testing, we found that the playback from the Bose speaker was the most similar to the original commands, *however they were still able to be reliably detected.*

### 7.3 Audio File Manipulation

Our adversary as defined in Section 4.1 is able to manipulate the audio that will be played over the speaker. Even with this capability, our technique will still function as intended. This is because our technique specifically targets artifacts produced by the speaker's physical design. An adversary could apply an equalizer, remove or amplify various ranges of frequency components, add in new frequency components, or any combination thereof and our technique will still work. Regardless of how the audio is manipulated before being played, the vibrations from playing any sound will cause the speaker's enclosure to resonant. This resonance is what allows our technique to detect the electronic speaker.

## 8 CONCLUSION

Voice interfaces have become an essential component of IoT devices used in many homes and offices. Unfortunately, the lack of command authentication has led to various injection attacks from different electronic speakers in its vicinity [21, 23]. These command injections have shown to have various consequences ranging from unauthorized used to financial exploit. To stop electronic speakers from injecting commands, we propose a detection mechanism based on the sub-bass over-excitation phenomena found in all speakers due to their enclosure casing. We demonstrate that our detection system can distinguish commands that originate from a human speaker from commands injected through an electronic speaker. By distinguishing between the two, we are able to prevent such attacks. In so doing, we dramatically reduce the ability of an attacker to inject potentially harmful commands to the voice assistants while having little effect on performance. To that end, we show that detection systems based on this phenomena significantly improve the security of voice assistants.

## 9 ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under grant number CNS-1702879. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## REFERENCES

[1] 2017 Voice Assistant Trends [Infographic]. https://ifttt.com/blog/2017/07/voice-

assistant-trends-infographic. 2017-07-12.
[2] Adobe demos "photoshop for audio," lets you edit speech as easily as text. https://arstechnica.com/information-technology/2016/11/adobe-voco-photoshop-for-audio-speech-editing/, 2017.
[3] Amazon Alexa Line. https://www.amazon.com/Amazon-Echo-And-Alexa-Devices/b?ie=UTF8&node=9818047011, 2017.
[4] Apple Siri. https://www.apple.com/ios/siri/, 2017.
[5] August Home Supports the Google Assistant. http://august.com/2017/03/28/google-assistant/, 2017.
[6] Cortana. https://www.microsoft.com/en-us/windows/cortana, 2017.
[7] Google Assistant. https://assistant.google.com/, 2017.
[8] Google Home now lets you shop by voice just like Amazon's Alexa. https://techcrunch.com/2017/02/16/google-home-now-lets-you-shop-by-voice-just-like-amazons-alexa/, 2017.
[9] Lyrebird. https://github.com/logant/Lyrebird, 2017.
[10] Starling Bank Integrates API into Google Home. http://bankinnovation.net/2017/02/starling-bank-integrates-api-into-google-home-video/, 2017.
[11] P. S. Aleksic and A. K. Katsaggelos. Audio-visual biometrics. *Proceedings of the IEEE*, 94(11):2025–2044, Nov 2006.
[12] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, Nov 2010.
[13] R. J. Baken and R. F. Orlikoff. *Clinical measurement of speech and voice*. Cengage Learning, 2000.
[14] L. Blue, H. Abdullah, L. Vargas, and P. Traynor. 2ma: Verifying voice commands via two microphone authentication. In *Proceedings of the 2018 ACM on Asia Conference on Computer and Communications Security*. ACM, 2018.
[15] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou. Hidden Voice Commands. In *25th USENIX Security Symposium*, 2016.
[16] G. Chetty and M. Wagner. LivenessâĂİ verification in audio-video authentication. 2004.
[17] N. Eveno and L. Besacier. Co-inertia analysis for "liveness" test in audio-visual biometrics. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, pages 257–261, Sept 2005.
[18] Google. Transactions Developer Preview. https://developers.google.com/actions/transactions/, 2017.
[19] A. K. Jain, R. Bolle, and S. Pankanti. *Biometrics: personal identification in networked society*, volume 479. Springer Science & Business Media, 2006.
[20] H. Kuwabara. Acoustic properties of phonemes in continuous speech for different speaking rate. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 4, pages 2435–2438. IEEE, 1996.
[21] S. Maheshwari. Burger King 'O.K. Google' Ad Doesn't Seem O.K. With Google. https://www.nytimes.com/2017/04/12/business/burger-king-tv-ad-google-home.html, 2017.
[22] D. Mukhopadhyay, M. Shirvanian, and N. Saxena. *All Your Voices are Belong to Us: Stealing Voices to Fool Humans and Machines*. 20th European Symposium on Research in Computer Security, 2015.
[23] S. Nichols. TV anchor says live on-air 'Alexa, order me a dollhouse' - Guess what happens next. https://www.theregister.co.uk/2017/01/07/tv-anchor-says-alexa-buy-me-a-dollhouse-and-she-does/, 2017.
[24] D. A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17(1):91 – 108, 1995.
[25] A. Ross and A. Jain. Information fusion in biometrics. *Pattern Recognition Letters*, 24(13):2115 – 2125, 2003. Audio- and Video-based Biometric Person Authentication (AVBPA 2001).
[26] R. R. Sanders. *The electrostatic loudspeaker design cookbook*. Audio Amateur, Incorporated, 2017.
[27] C. Sanderson and K. K. Paliwal. Identity verification using speech and face information. *Digital Signal Processing*, 14(5):449 – 480, 2004.
[28] M. Shirvanian and N. Saxena. Wiretapping via Mimicry: Short Voice Imitation Man-in-the-middle attacks on Crypto Phones. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 2014.
[29] S. T. Shivappa, M. M. Trivedi, and B. D. Rao. Audiovisual information fusion in human 2013;computer interfaces and intelligent environments: A survey. *Proceedings of the IEEE*, 98(10):1692–1715, Oct 2010.
[30] H. Stephenson. UX design trends 2018: from voice interfaces to a need to not trick people. Digital Arts - https://www.digitalartsonline.co.uk/features/interactive-design/ux-design-trends-2018-from-voice-interfaces-need-not-trick-people/, 2018.
[31] K. N. Stevens. *Acoustic phonetics*, volume 30. MIT press, 2000.
[32] S. Studio. Respeaker 4-mic array for raspberry pi. https://www.seeedstudio.com/ReSpeaker-4-Mic-Array-for-Raspberry-Pi-p-2941.html. Accessed: March 5, 2018.
[33] A. Team. Audacity homepage. https://www.audacityteam.org/. Accessed: March 5, 2018.
[34] I. R. Titze and D. W. Martin. *Principles of voice production*. ASA, 1998.
[35] T. Vaidya, Y. Zhang, M. Sherr, and C. Shields. Cocaine Noodles: Exploiting the Gap Between Human and Machine Speech Recognition. *11th USENIX Workshop on Offensive Technologies*, 2015.
[36] Z. Wu, S. Gao, E. S. Cling, and H. Li. A study on replay attack and anti-spoofing for text-dependent speaker verification. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pages 1–5, Dec 2014.
[37] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu. Dolphinattack: Inaudible Voice Commands. *Computer and Communications Security (CCS)*, 2017.

## A COMMAND PHRASES

For our experiments we used the following eight command phrases. These were picked to simulate real queries to voice interfaces and forced the organic speaker to enunciate each sentence. The last entry is explicitly used to force the speaker to voice a variety of phonemes.

(1) "O.K. Google, Browse to evil.com."
(2) "O.K. Google, call grandma."
(3) "O.K. Google, record a video."
(4) "Hey Google, text John buy spam today."
(5) "Hey Google, post I'm so evil on Twitter."
(6) "Alexa, call grandpa."
(7) "Alexa, text mom what was my social security number again?"
(8) "These puffy tarantulas cross bravely shepherding homeless grouper through explosions."

## B ELECTRONIC SPEAKERS


**Figure 10: We ran our experiments using a wide range of speakers that vary in quality; A: Music Angel (MSRP: 2014 - $16.99), B: HP Computer Speakers (MSRP: 2009 - $50.00), C: Dynex TV (MSRP: 2010 - $250), D: Acer Predator monitor speakers (MSRP: 2018 - $399), E: Samsung Sound Bar (MSRP: 2018 - $280), F: Insigna TV (MSRP: 2017 - $330), G: MacBook Air (MSRP: 2015 - $1000), and H: Bose Wave IV (MSRP: 2017 - $499)**