Inaudible Voice Commands: The Long-Range Attack and Defense

Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury,
University of Illinois at Urbana-Champaign

https://www.usenix.org/conference/nsdi18/presentation/roy

This paper is included in the Proceedings of the
15th USENIX Symposium on Networked
Systems Design and Implementation (NSDI '18).
April 9–11, 2018 • Renton, WA, USA
ISBN 978-1-939133-01-4

Inaudible Voice Commands: The Long-Range Attack and Defense

Nirupam Roy, Sheng Shen, Haitham Hassanieh, Romit Roy Choudhury
University of Illinois at Urbana-Champaign

Abstract

Recent work has shown that inaudible signals (at ultrasound frequencies) can be designed in a way that they become audible to microphones. Designed well, this can empower an adversary to stand on the road and silently control Amazon Echo and Google Home-like devices in people's homes. A voice command like "Alexa, open the garage door" can be a serious threat.

While recent work has demonstrated feasibility, two issues remain open: (1) The attacks can only be launched from within $5ft$ of Amazon Echo, and increasing this range makes the attack audible. (2) There is no clear solution against these ultrasound attacks, since they exploit a recently discovered loophole in hardware non-linearity.

This paper is an attempt to close both these gaps. We begin by developing an attack that achieves $25ft$ range, limited by the power of our amplifier. We then develop a defense against this class of voice attacks that exploit non-linearity. Our core ideas emerge from a careful forensics on voice, i.e., finding indelible traces of non-linearity in recorded voice signals. Our system, *LipRead*, demonstrates the inaudible attack in various conditions, followed by defenses that only require software changes to the microphone.

1 Introduction

A number of recent research papers have focused on the topic of inaudible voice commands [37, 48, 39]. Backdoor [37] showed how hardware non-linearities in microphones can be exploited, such that *inaudible ultrasound signals* can become audible to any microphone. DolphinAttack [48] developed on Backdoor to demonstrate that no software is needed at the microphone, i.e., a voice enabled device like Amazon Echo can be made to respond to inaudible voice commands. A similar paper independently emerged in arXiv [39], with a video demonstration of such an attack [3]. These attacks are becoming increasingly relevant, particularly with the proliferation of voice enabled devices including Amazon Echo, Google Home, Apple Home Pod, Samsung refrigerators, etc.

While creative and exciting, these attacks are still deficient on an important parameter: *range*. DolphinAttack can launch from a distance of $5ft$ to Amazon Echo [48] while the attack in [39] achieves $10ft$ by becoming partially audible. In attempting to enhance range, we realized strong tradeoffs with inaudibility, i.e., the output of the speaker no longer remains silent. This implies that currently known attacks are viable in short ranges, such as Alice's friend visiting Alice's home and silently attacking her Amazon Echo [11, 48]. However, the general, and perhaps more alarming attack, is the one in which the attacker parks his car on the road and controls voice-enabled devices in the neighborhood, and even a person standing next to him does not hear it. This paper is an attempt to achieve such an attack radius, followed by defenses against them. We formulate the core problem next and outline our intuitions and techniques for solving them.

Briefly, non-linearity is a hardware property that makes high frequency signals arriving at a microphone, say s_{hi}, get shifted to lower frequencies s_{low} (see Figure 1). If s_{hi} is designed carefully, then s_{low} can be almost identical to s_{hi} but shifted to within the audibility cutoff of $20kHz$ inside the microphone. As a result, even though humans do not hear s_{hi}, non-linearity in microphones produces s_{low}, which then become legitimate voice commands to devices like Amazon Echo. This is the root opportunity that empowers today's attacks.

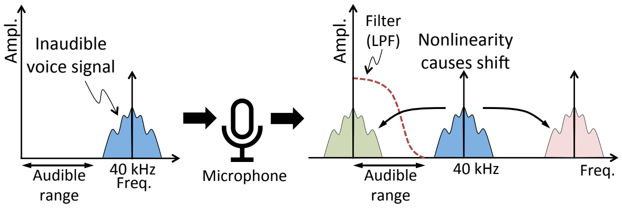


Figure 1: Hardware non-linearity creates frequency shift. Voice commands transmitted over inaudible ultrasound frequencies get shifted into the lower audible bands after passing through the non-linear microphone hardware.

Two important points need mention at this point. (1) Non-linearity triggers at high frequencies and at high power – if s_{hi} is a soft signal, then the non-linear effects do not surface. (2) Non-linearity is fundamental to acoustic hardware and is equally present in speakers as in microphones. Thus, when s_{hi} is played through speakers, it will also undergo the frequency shift, producing an audible s_{low}. Dolphin and other attacks sidestep this problem by operating at low power, thereby forcing the output of the speaker to be almost inaudible. This inherently limits the range of the attack to $5ft$; any attempt to increase this range will result in audibility.

This paper breaks away from the zero sum game between range and audibility by an alternative transmitter design. Our core idea is to use multiple speakers, and stripe segments of the voice signal across them such that leakage from each speaker is narrow band, and confined to low frequencies. This still produces a garbled, audible sound. To achieve true inaudibility, we solve a min-max optimization problem on the length of the voice segments. The optimization picks the segment lengths in a way such that the aggregate leakage function is completely below the human auditory response curve (i.e., the minimum separation between the leakage and the human audibility curve is maximized). This ensures, by design, the attack is inaudible.

Defending against this class of non-linearity attacks is not difficult if one were to assume hardware changes to the receiver (e.g., Amazon Echo or Google Home). An additional ultrasound microphone will suffice since it can detect the s_{hi} signals in air. However, with software changes alone, the problem becomes a question of forensics, i.e., can the shifted signal s_{low} be discriminated from the same legitimate voice command, s_{leg}. In other words, does non-linearity leave an indelible trace on s_{low} that would otherwise not be present in s_{leg}.

Our defense relies on the observation that voice signals exhibit well-understood structure, composed of fundamental frequencies and harmonics. When this structure passes through non-linearity, part of it remains preserved in the shifted and blended low frequency signals. In contrast, legitimate human voice projects almost no energy in these low frequency bands. An attacker that injects distortion to hide the traces of voice, either pollutes the core voice command, or raises the energy floor in these bands. This forces the attacker into a zero-sum game, disallowing him from erasing the traces of non-linearity without raising suspicion.

Our measurements confirm the possibility to detect voice traces, i.e., even though non-linearity superimposes many harmonics and noise signals on top of each other, and attenuates them significantly, cross-correlation still reveals the latent voice fingerprint. Of course, various intermediate steps of contour tracking, filtering, frequency-selective compensation, and phoneme correlation are necessary to extract out the evidence. Nonetheless, our final classifier is transparent and does not require any training at all, but succeeds for voice signals only, as opposed to the general class of inaudible microphone attacks (such as jamming [37]). We leave this broader problem to future work.

Our overall system *LipRead* is built on multiple platforms. For the inaudible attack at long ranges, we have developed an ultrasound speaker array powered by our custom-made amplifier. The attacker types a command on the laptop, MATLAB converts the command to a voice signal, and the laptop sends this through our amplifier to the speaker. We demonstrate controlling Amazon Echo, iPhone Siri, and Samsung devices from a distance of $25ft$, limited by the power of our amplifier. For defense, we record signals from Android Samsung S6 phones, as well as from off-the-shelf microphone chips (popular in today's devices). We attack the system with various ultrasound commands, both from literature as well as our own. *LipRead* demonstrates defense against all attacks with 97% precision and 98% recall. The performance remains robust across varying parameters, including multipath, power, attack location, and various signal manipulations.

Current limitations: Our long-range attacks have been launched from within a large room, or from outside a house with open windows. When doors and windows were closed, the attack was unsuccessful since our high-frequency signals attenuated while passing through the wall/glass. We believe this is a function of power, however, a deeper treatment is necessary around this question. In particular: (1) Will high power amplifiers be powerful enough for high-frequency signals to penetrate such barriers? (2) Will high-power and high-frequency signals trigger non-linearity inside human ears? (3) Are there other leakages that will emerge in such high power and high frequency regimes. We leave these questions to future work.

In sum, our core contributions may be summarized as follows:

- A transmitter design that breaks away from the tradeoff between attack range and audibility. The core ideas pertain to carefully striping frequency bands across an array of speakers, such that individual speakers are silent but the microphone is activated.

- A defense that identifies human voice traces at very low frequencies (where such traces should not be present) and uses them to protect against attacks that attempt to erase or disturb these traces.

The subsequent sections elaborate on these ideas, beginning with some relevant background on non-linearity, followed by threat model, attack design, and defense.

2 Background: Acoustic Non-linearity

Microphones and speakers are in general designed to be linear systems, meaning that the output signals are linear combinations of the input. In the case of power amplifiers inside microphones and speakers, if the input sound signal is $s(t)$, then the output should ideally be:

$$s_{out}(t) = A_1 s(t)$$

where A_1 is the amplifier gain. In practice, however, acoustic components in microphones and speakers (like diaphragms, amplifiers, etc.) are linear only in the audible frequency range ($< 20kHz$). In ultrasound bands ($> 25kHz$), the responses exhibit non-linearity [28, 19, 16, 38, 22]. Thus, for ultrasound signals, the output of the amplifier becomes:

$$\begin{aligned} s_{out}(t) = & \sum_{i=1}^{\infty} A_i s^i(t) = A_1 s(t) + A_2 s^2(t) + A_3 s^3(t) + \ldots \\ \approx & \ A_1 s(t) + A_2 s^2(t) \end{aligned} \tag{1}$$

Higher order terms are typically extremely weak since $A_{4+} \ll A_3 \ll A_2$ and hence can be ignored.

Recent work [37] has shown ways to exploit this phenomenon, i.e., it is possible to play ultrasound signals that cannot be heard by humans but can be directly recorded by any microphone. Specifically, an ultrasound speaker can play two inaudible tones: $s_1(t) = \cos(2\pi f_1 t)$ at frequency $f_1 = 38$kHz and $s_2 = \cos(2\pi f_2 t)$ at frequency $f_2 = 40$kHz. Once the combined signal $s_{hi}(t) = s_1(t) + s_2(t)$ passes through the microphone's nonlinear hardware, the output becomes:

$$\begin{aligned} s_{out}(t) = & \ A_1 s_{hi}(t) + A_2 s_{hi}^2(t) \\ = & \ A_1(s_1(t) + s_2(t)) + A_2(s_1(t) + s_2(t))^2 \\ = & \ A_1 \cos(2\pi f_1 t) + A_1 \cos(2\pi f_2 t) \\ + & \ A_2 \cos^2(2\pi f_1 t) + A_2 \cos^2(2\pi f_2 t) \\ + & \ 2A_2 \cos(2\pi f_1 t) \cos(2\pi f_2 t) \end{aligned}$$

The above signal has frequency components at f_1, f_2, $2f_1$, $2f_2$, $f_2 + f_1$, and $f_2 - f_1$. This can be seen by expanding the equation:

$$\begin{aligned} s_{out}(t) = & \ A_1 \cos(2\pi f_1 t) + A_1 \cos(2\pi f_2 t) \\ + & \ A_2 + 0.5A_2 \cos(2\pi 2 f_1 t) + 0.5A_2 \cos(2\pi 2 f_2 t) \\ + & \ A_2 \cos(2\pi(f_1 + f_2)t) + A_2 \cos(2\pi(f_2 - f_1)t) \end{aligned}$$

Before digitizing and recording the signal, the microphone applies a low pass filter to remove frequency components above the microphone's cutoff of $24KHz$. Observe that f_1, f_2, $2f_1$, $2f_2$, and $f_1 + f_2$ are all > 24kHz. Hence, what remains (as acceptable signal) is:

$$s_{low}(t) = A_2 + A_2 \cos(2\pi(f_2 - f_1)t) \tag{2}$$

This is essentially a $f_2 - f_1 = 2$kHz tone which will be recorded by the microphone. However, this demonstrates the core opportunity, i.e., by sending a *completely inaudible signal*, we are able to generate an audible "copy" of it inside any unmodified off-the-shelf microphone.

3 Inaudible Voice Attack

We begin by explaining how the above non-linearity can be exploited to send inaudible commands to *voice enabled devices* (VEDs) at a short range. We identify deficiencies in such an attack and then design the longer range, truly inaudible attack.

3.1 Short Range Attack

Let $v(t)$ be a baseband voice signal that once decoded translates to the command: "Alexa, mute yourself". An attacker moves this baseband signal to a high frequency $f_{hi} = 40$kHz (by modulating a carrier signal), and plays it through an ultrasound speaker. The attacker also plays a tone at $f_{hi} = 40$kHz. The played signal is:

$$s_{hi}(t) = \cos(2\pi f_{hi} t) + v(t) \cos(2\pi f_{hi} t) \tag{3}$$

After this signal passes through the non-linear hardware and low-pass filter of the microphone, the microphone will record:

$$s_{low}(t) = \frac{A_2}{2} \left(1 + v^2(t) + 2v(t)\right) \tag{4}$$

This shifted signal contains a strong component of $v(t)$ (due to more power in the speech components), and hence, gets decoded correctly by almost all microphones.

■ **What happens to $v^2(t)$?**

Figure 2 shows the power spectrum $V(f)$ corresponding to the voice command $v(t) =$"Alexa, mute yourself". Here the power spectrum corresponding to $v^2(t)$ which is equal to $V(f) * V(f)$ where $(*)$ is the convolution operation. Observe that the spectrum of the human voice is between $[50 - 8000]Hz$ and the relatively weak components of $v^2(t)$ line up underneath the voice frequencies after convolution. A component of $v^2(t)$ also falls at DC, however, degrades sharply. The overall weak presence of $v^2(t)$ leaves the $v(t)$ signal mostly unharmed, allowing VEDs to decode the command correctly.

However, to help $v(t)$ enter the microphone through the "non-linear inlet", $s_{hi}(t)$ must be transmitted at sufficiently high power. Otherwise, $s_{low}(t)$ will be buried in noise (due to small A_2). *Unfortunately, increasing the transmit power at the speaker triggers non-linearities at the speaker's own diaphragm and amplifier*, resulting in an audible $s_{low}(t)$ at the output of the speaker. Since $s_{low}(t)$ contains the voice command $v(t)$, the attack becomes audible. Past attacks sidestep this problem by operating at low power, thereby forcing the output of the

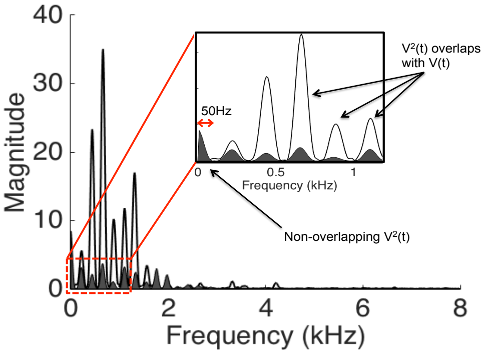


Figure 2: Spectrum of $V(f) * V(f)$ which is the non-linear leakage after passing through the microphone

speaker to be almost inaudible [49]. This inherently limits the radius of attack to a short range of $5ft$. Attempts to increase this range results in audibility, defeating the purpose of the attack.

Figure 3 confirms this with experiments in our building. Five volunteers visited marked locations and recorded their perceived loudness of the speaker's leakage. Clearly, speaker non-linearity produces audibility, a key problem for long range attacks.

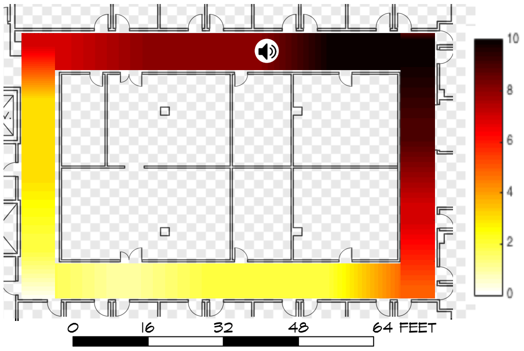


Figure 3: Heatmap showing locations at which $v(t)$ leakage from the speaker is audible.

3.2 Long Range Attack

Before developing the long range attack, we concisely present the assumptions and constraints on the attacker.

■ **Threat Model:** We assume that:

- The attacker cannot enter the home to launch the attack, otherwise, the above short range attack suffices.
- The attacker cannot leak any audible signals (even in a beamformed manner), otherwise such inaudible attacks are not needed in the first place.
- The attacker is resourceful in terms of hardware and energy (perhaps the attacking speaker can be carried in his car or placed at his balcony, pointed at VEDs in surrounding apartments or pedestrians).
- In case the receiver device (e.g., Google Home) is voice fingerprinted, we assume the attacker can synthesize the legitimate user's voice signal using known techniques [46, 5] to launch the attack.
- The attacker cannot estimate the precise *channel impulse response* (CIR) from its speaker to the voice enabled device (VED) that it intends to attack.

■ **Core Attack Method:**
LipRead develops a new speaker design that facilitates considerably longer attack range, while eliminating the audible leakage at the speaker. Instead of using one ultrasound speaker, *LipRead* uses multiple of them, physically separated in space. Then, *LipRead* splices the spectrum of the voice command $V(f)$ into carefully selected segments and plays each segment on a different speaker, thereby limiting the leakage from each speaker.

■ **The Need for Multiple Speakers:**
To better understand the motivation, let us first consider using two ultrasound speakers. Instead of playing $s_{hi}(t) = \cos(2\pi f_{hi}t) + v(t)\cos(2\pi f_{hi}t)$ on one speaker, we now play $s_1(t) = \cos(2\pi f_{hi}t)$ on the first speaker and $s_2(t) = v(t)\cos(2\pi f_{hi}t)$ on the second speaker where $f_{hi} = 40kHz$. In this case, the 2 speakers will output:

$$\begin{aligned} s_{out1} &= \cos(2\pi f_{hi}t) + \cos^2(2\pi f_{hi}t) \\ s_{out2} &= v(t)\cos(2\pi f_{hi}t) + v^2(t)\cos^2(2\pi f_{hi}t) \end{aligned} \tag{5}$$

For simplicity, we ignore the terms A_1 and A_2 (since they do not affect our understanding of frequency components). Thus, when s_{out1} and s_{out2} emerge from the two speakers, human ears filter out all frequencies $> 20kHz$. What remains audible is only:

$$\begin{aligned} s_{low1} &= 1/2 \\ s_{low2} &= v^2(t)/2 \end{aligned}$$

Observe that neither s_{low1} nor s_{low2} contains the voice signal $v(t)$, hence the actual attack command is no longer audible with two speakers. However, the microphone under attack will still receive the aggregate ultrasound signal from the two speakers, $s_{hi}(t) = s_1(t) + s_2(t)$, and its own non-linearity will cause a "copy" of $v(t)$ to get shifted into the audible range (recall Equation 4). Thus, this 2-speaker attack activates VEDs from greater distances, while the actual voice command remains inaudible to bystanders.

Although the voice signal $v(t)$ is inaudible, signal $v^2(t)$ still leaks and becomes audible (especially at higher power). This undermines the attack.

■ **Suppressing $v^2(t)$ Leakage:**
To suppress the audibility of $v^2(t)$, *LipRead* expands to N ultrasound speakers. It first partitions the audio spectrum

$V(f)$ of the command signal $v(t)$, ranging from f_0 to f_N, into N frequency bins: $[f_0, f_1]$, $[f_1, f_2]$, $\cdots$, $[f_{N-1}, f_N]$ as shown in Fig. 4. This can be achieved by computing an FFT of the signal $v(t)$ to obtain $V(f)$. $V(f)$ is then multiplied with a rectangle function $rect(f_i, f_{i+1})$ which gives a filtered $V_{[f_i, f_{i+1}]}(f)$. An IFFT is then used to generate $v_{[f_i, f_{i+1}]}(t)$ which is multiplied by an ultrasound tone $\cos(2\pi f_{hi}t)$ and outputted on the i^{th} ultrasound speaker as shown in Fig. 4.

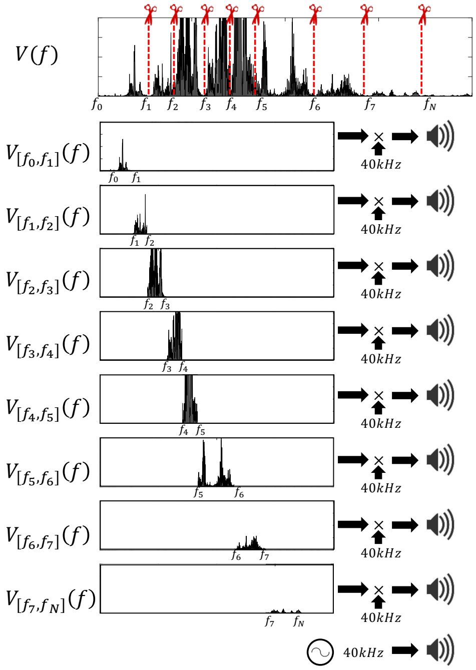


Figure 4: Spectrum Splicing: optimally segmenting the voice command frequencies and playing it through separate speakers so that the net speaker-output is silent.

In this case, the audible leakage from i^{th} ultrasound speaker will be $s_{low,i}(t) = v^2_{[f_i, f_{i+1}]}(t)$. In the frequency domain, we can write this leakage as:

$$S_{low,i}(f) = \quad V_{[f_i,f_{i+1}]}(f) * V_{[f_i,f_{i+1}]}(f)$$

This leakage has two important properties:

$$\begin{aligned} &(1) \quad E\left[|S_{low,i}(f)|^2\right] \leq E\left[|V(f) * V(f)|^2\right] \\ &(2) \quad BW(S_{low,i}(f)) \leq BW(V(f) * V(f)) \end{aligned}$$

where $E[|.|^2]$ is the power of audible leakage and $BW(.)$ is the bandwidth of the audible leakage due to non-linearities at each speaker. The above properties imply that splicing the spectrum into multiple speakers reduces the audible leakage from any given speaker. It also reduces the bandwidth and hence concentrates the audible leakage in a smaller band below 50 Hz.

While per-speaker leakage is smaller, they can still add up to become audible. The total leakage power can be written as:

$$L(f) = \left| \sum_{i=1}^{N} V_{[f_i,f_{i+1}]}(f) * V_{[f_i,f_{i+1}]}(f) \right|^2$$

To achieve true inaudibility, we need to ensure that the total leakage is not audible. To address this challenge, we leverage the fact that humans cannot hear the sound if the sound intensity falls below certain threshold, which is frequency dependent. This is known as the “Threshold of Hearing Curve”, $T(f)$. Fig. 5 shows $T(f)$ in dB as function of frequency. Any sound with intensity below the threshold of hearing will be inaudible.

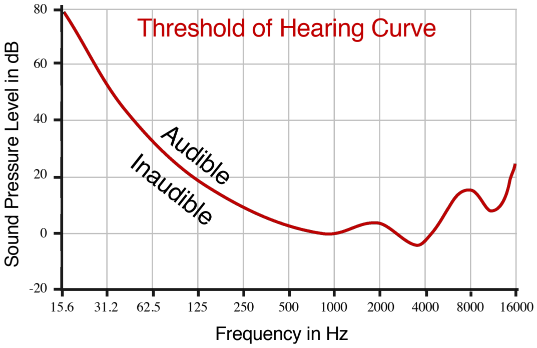


Figure 5: Threshold of Hearing Curve

LipRead aims to push the total leakage spectrum, $L(f)$, below the “Threshold of Hearing Curve” $T(f)$. To this end, *LipRead* finds the best partitioning of the spectrum, such that the leakage is below the threshold of hearing. If multiple partitions satisfy this constraint, *LipRead* picks the one that has the largest gap from the threshold of hearing curve. Formally, we solve the below optimization problem:

$$\begin{aligned} &\underset{\{f_1, f_2, \cdots, f_{N-1}\}}{\text{maximize}} \; \min_{f}[T(f) - L(f)] \\ &\text{subject to} f_0 \leq f_1 \leq f_2 \leq \cdots \leq f_N \end{aligned} \tag{6}$$

The solution partitions the frequency spectrum to ensure that the leakage energy is below the hearing threshold for every frequency bin. This ensures inaudibility at any human ear.

■ Increasing Attack Range:
It should be possible to increase attack range with more speakers, while also limiting audible leakage below the required hearing threshold. This holds in principle due to the following reason. For a desired attack range, say r, we can compute the minimum power density (i.e., power per frequency) necessary to invoke the VED. This power P_r needs to be high since the non-linear channel will strongly attenuate it by the factor A_2. Now consider the worst case where a voice command has equal magnitude in all frequencies. Given each frequency needs power P_r and each speaker’s output needs to be below *threshold*

of hearing for all frequencies, we can run our *min-max optimization* for increasing values of N, where N is the number of speakers. The minimum N that gives a feasible solution is the answer. Of course, this is the upper bound; for a specific voice signal, N will be lower.

Increasing speakers can be viewed as beamforming the energy towards the VED. In the extreme case for example, every speaker will play one frequency tone, resulting in a strong DC component at the speaker's output which would still be inaudible. In practice, our experiments are bottlenecked by ADCs, amplifiers, speakers, etc., hence we will report results with an array of 61 small ultrasound speakers.

4 Defending Inaudible Voice Commands

Recognizing inaudible voice attacks is essentially a problem of acoustic forensics, i.e., detecting evidence of nonlinearity in the signal received at the microphone. Of course, we assume the attacker knows our defense techniques and hence will try to remove any such evidence. Thus, the core question comes down to: *is there any trace of non-linearity that just cannot be removed or masked?*

To quantify this, let $v(t)$ denote a human voice command signal, say "Alexa, mute yourself". When a human issues this command, the recorded signal $s_{leg} = v(t) + n(t)$, where $n(t)$ is noise from the microphone. When an attacker plays this signal over ultrasound (to launch the non-linear attack), the recorded signal s_{nl} is:

$$s_{nl} = \frac{A_2}{2}(1 + 2v(t) + v^2(t)) + n(t) \tag{7}$$

Figure 6 shows an example of s_{leg} and s_{nl}. Evidently, both are very similar, and both invoke the same response in VEDs (i.e., the text-to-speech converter outputs the same text for both s_{leg} and s_{nl}). A defense mechanism would need to examine any incoming signal s and tell if it is low-frequency legitimate or a shifted copy of the high-frequency attack.

4.1 Failed Defenses

Before we describe *LipRead*'s defense, we mention a few other possible defenses which we have explored before converging on our final defense system. We concisely summarize 4 of these ideas.

■ Decompose Incoming Signal $s(t)$:
One solution is to solve for $s(t) = \frac{A_2}{2}(1 + 2\hat{v}(t) + \hat{v}^2(t))$, and test if the resulting $\hat{v}(t)$ produces the same text-to-speech (T2S) output as $s(t)$. However, this proved to be a fallacious argument because, if such a $\hat{v}(t)$ exists, it will always produce the same T2S output as $s(t)$. This is because such a $\hat{v}(t)$ would be a cleaner version of the voice command (without the non-linear component); if the polluted version s passes the T2S test, the cleaner version obviously will.

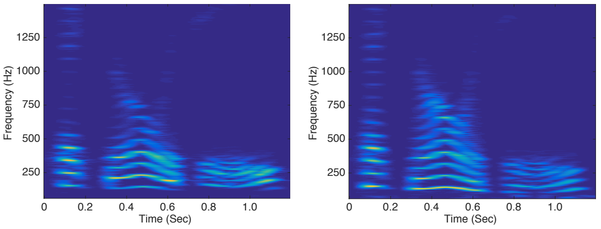


Figure 6: Spectrogram for s_{leg} and s_{nl} for voice command "Alexa, mute yourself".

■ Energy at Low Frequencies from $v^2(t)$:
Another solution is to extract portions of $s(t)$ from the lower frequencies – since regular voice signals do not contain sub-$50Hz$ components, energy detection should offer evidence. Unfortunately, environmental noise (e.g., fans, A/C machines, wind) leaves non-marginal residue in these low bands. Moreover, an attacker could deliberately reduce the power of its signal so that its leakage into sub-$50Hz$ is small. Our experiments showed non-marginal false positives in the presence of environmental sound and soft attack signals.

■ Amplitude Degradation at Higher Frequencies:
The air absorbs ultrasound frequencies far more than voice (which translates to sharper reduction in amplitude as the ultrasound signal propagates). Measured across different microphones separated by $\approx 7.3cm$ in Amazon Echo and Google Home, the amplitude difference should be far greater for ultrasound. We designed a defense that utilized the maximum amplitude slope between microphone pairs – this proved to be a robust discriminator between s_{leg} and s_{nl}. However, we were also able to point two (reasonably synchronized) ultrasound beams from opposite directions. This reduced the amplitude gradient, making it comparable to legitimate voice signals (Alexa treated the signals as multipath). In the real-world, we envisioned 2 attackers launching this attack by standing at 2 opposite sides of a house. Finally, this solution would require an array of microphones on the voice enabled device. Hence, it is inapplicable to one or two microphone systems (like phones, wearables, refrigerators).

■ Phase Based Separation of Speakers:
Given that long range attacks need to use at least 2 speakers (to bypass speaker non-linearity), we designed an angle-of-arrival (AoA) based technique to estimate the physical separation of speakers. In comparison to human voice, the source separation consistently showed success, so long as the speakers are more than $2cm$ apart. While practical attacks would certainly require multiple speakers, easily making them $2cm$ apart, we aimed at solving

the short range attack as well (i.e., where the attack is launched from a single speaker). Put differently, the right evidence of non-linearity should be one that is present regardless of the number of speakers used.

4.2 LipRead Defense Design

Our final defense is to search for traces of $v^2(t)$ in sub-50Hz. However, we now focus on exploiting the structure of human voice. The core observation is simple: voice signals exhibit well-understood patterns of fundamental frequencies, added to multiple higher order harmonics (see Figure 6). We expect this structure to partly reflect in the sub-50Hz band of $s(t)$ (that contains $v^2(t)$), and hence correlate with carefully extracted spectrum above-50Hz (which contains the dominant $v(t)$). With appropriate signal scrubbing, we expect the correlation to emerge reliably, however, if the attacker attempts to disrupt correlation by injecting sub-50Hz noise, the stronger energy in this low band should give away the attack. We intend to force the attacker into this zero sum game.

■ Key Question: Why Should $v^2(t)$ Correlate?

Figure 7(a) shows a simplified abstraction of a legitimate voice spectrum, with a narrow fundamental frequency band around f_j and harmonics at integer multiples nf_j. The lower bound on f_j is $> 50Hz$ [41]. Now recall that when this voice spectrum undergoes non-linearity, each of f_j and nf_j will self-convolve to produce "copies" of themselves around DC (Figure 7(b)). Of course, the A_2 term from non-linearity strongly attenuates this "copy". *However, given the fundamental band around f_j and the harmonics around nf_j are very similar in structure, each of $\approx 20Hz$ bandwidth, the energy between $[0, 20kHz]$ superimposes.* This can be expressed as:

$$E_{[0,20]} \approx E\left[A_2 \sum_{n=1}^{N} |V_{[nf_j-20,\ nf_j+20]} * V_{[nf_j-20,\ nf_j+20]}|^2\right] \tag{8}$$

The net result is distinct traces of energy in sub-20Hz bands, and importantly, this energy variation (over time) mimics that of f_j. For a legitimate attack, on the other hand, the sub-20Hz is dominantly uncorrelated hardware and environmental noise.

Figure 8(a) and (b) zoom into sub-50Hz and compare the traces of energy for s_{leg} and s_{nl}, respectively. The s_{nl} signal clearly shows more energy concentration, particularly when the actual voice signal is strong. Figure 9 plots the power in the sub-50Hz band with increasing voice loudness levels for both s_{leg} and s_{nl}. Note that loudness level is expressed in $dBSpl$, where Spl denotes "sound pressure level", the standard metric for measuring sound. Evidently, non-linearity shows increasing power due to the self-convolved spectrum overlapping in the lower band. Legitimate voice signals generate significantly less energy in these bands, thereby remaining flat for higher loudness.

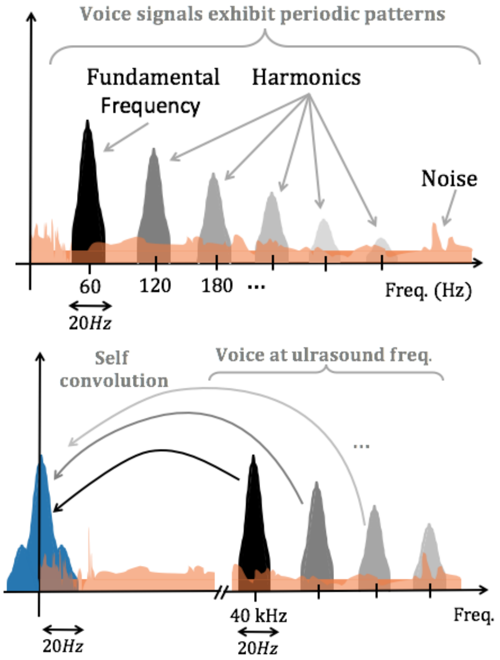
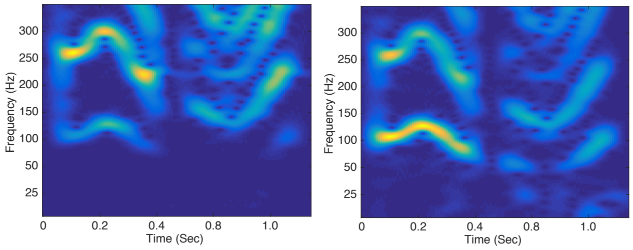


Figure 7: (a) A simplified voice spectrum showing the structure. (b) Voice spectra after non-linear attack.



(a) (b)

Figure 8: Spectrogram of the (a) audible and (b) inaudible attack voice. The attack signal contains higher power below 50Hz, indicated by lighter color.

■ Correlation Design

The width of the fundamental frequencies and harmonics are time-varying, however, at any given time, if it is $B\ Hz$, then the self-convolved signal gets shifted into $[0, B]Hz$ as well. Note that this is independent of the actual values of center frequencies, f_j and nf_j. Now, let $s_{<B}(t)$ denote the sub-$B\ Hz$ signal received by the microphone and $s_{>B}(t)$ be the signal above $B\ Hz$ that contains the voice command. *LipRead* seeks to correlate the energy variation over time in $s_{<B}(t)$ with the energy variation at the fundamental frequency, f_j in $s_{>B}(t)$. We track the fundamental frequency in $s_{>B}(t)$ using standard acoustic libraries, but then average the power around $B\ Hz$ of this frequency. This produces a power profile over time, P_{f_j}. For $s_{<B}(t)$, we also track the average power

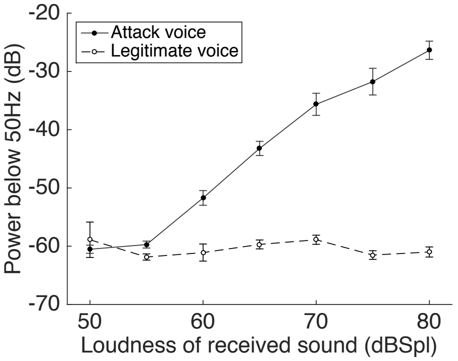


Figure 9: The loudness vs sub-50Hz band power plot.

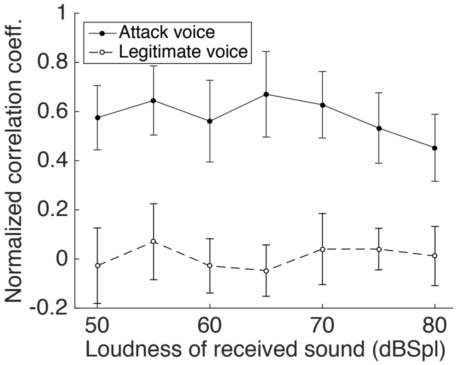


Figure 10: The loudness vs correlation between P_{f_j} and $s_{<B}(t)$, denoting the power variation of the fundamental frequency and the sub 20Hz band, respectively.

over time. However, to avoid weak signals and disruption from noise, we remove time windows in which f_j's power is below its average. We stitch together the remaining windows from both P_{f_j} and $s_{<B}(t)$ and compute their correlation co-efficient. We use an average value of $B = 20Hz$.

Figure 10 shows the correlation for increasing loudness levels of the recorded signal (loudness below 60$dBSpl$ is not audible). The comparison is against a legitimate voice command. Evidently, we recorded consistent correlation gap, implying that non-linearity is leaving some trace in the low-frequency bands, and this trace preserves some structure of the actual voice signal. Of course, we have not yet accounted for the possibility that the attacker can inject noise to disrupt correlation.

■ Improved Attack via Signal Shaping

The natural question for the attacker is how to modify/add signals such that the correlation gap gets narrowed. Several possibilities arise:

(1) Signal $-v^2(t)$ can be added to the speaker in the low frequency band and transmitted with the high frequency ultrasound $v(t)$. Given that ultrasound will produce $v^2(t)$ after non-linearity, and $-v^2(t)$ will remain as is, the two should interact at the microphone and cancel. Unfortunately, channels for low frequencies and ultrasound are different and unknown, hence it is almost impossible to design the precise $-v^2(t)$ signal. Of course, we will still attempt to attack with such a deliberately shaped signal.

(2) Assuming the ultrasound $v(t)$ has been up-converted to $[40,\ 44]kHz$, the attacker could potentially concatenate spurious frequencies from say $[44,\ 46]kHz$. These frequencies would also self-convolve and get "copied" around DC. This certainly affects correlation since these spurious frequencies would not correlate well (in fact, they can be designed to not correlate). The attacker's hope should be to lower correlation while maintaining a low energy footprint below 20Hz.

The attacker can use the above approaches to try to defeat the zero-sum game. Figure 11 plots results from 4000 attempts to achieve low correlation and low energy. Of these, 3500 are random noises injected in legitimate voice commands, while the remaining 500 are more carefully designed distortions (such as frequency concatenation, phase distortions, low frequency injection, etc.). Of course, in all these cases, the distorted signal was still correct, i.e., the VED device responded as it should.

On the other hand, 450 different legitimate words were spoken by different humans (shown as hollow dots), at various loudness levels, and accents, and styles. Clusters emerge suggesting promise of separation. However, some commands were still too close, implying the need for greater margin of separation.

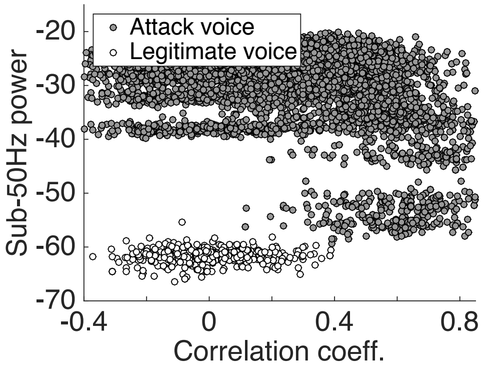
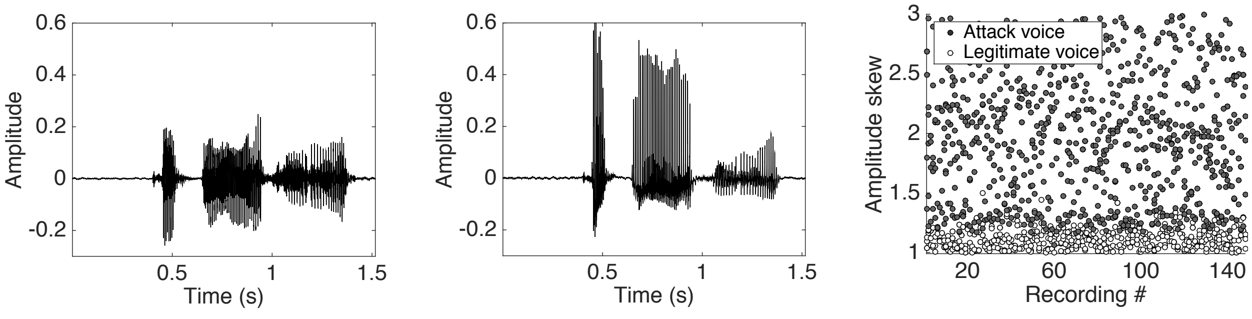


Figure 11: Zero sum game between correlation and power at sub-50Hz bands. Attacker attempts to reduce correlation by signal shaping or noise injection at sub-50Hz band.

■ Leveraging Amplitude Skew from $v^2(t)$

In order to increase the separation margin, *LipRead* leverages the amplitude skew resulting from $v^2(t)$. Specifically, two observations emerge: (1) When the harmonics in voice signals self-convolve to form $v^2(t)$, they fall at the same frequencies of the harmonics (since the gaps between the harmonics are quite homogeneous). (2) The signal $v^2(t)$ is a time domain signal with only posi-



(a) (b) (c)

Figure 12: (a) Sound signals in time domain from s_{leg} and (b) s_{nl}, demonstrating a case of amplitude skew. (c) Amplitude skew for various attack and legitimate voice commands.

tive amplitude. Combining these together, we postulated that amplitudes of the harmonics would be positively biased, especially for those that are strong (since $v^2(t)$ will be relatively stronger at that location). In contrast, amplitudes of legitimate voice signals should be well balanced on the positive and negative. Figure 12(a,b) shows one contrast between a legitimate voice s_{leg} and the recorded attack signal s_{nl}. In pursuit of this opportunity, we extract the ratio of the maximum and minimum amplitude (we average over the top 10% for robustness against outliers). Using this as the third dimension for separation, Figure 12(c) re-plots the s_{leg} and s_{nl} clusters. While the separation margin is close, combining it with correlation and power, the separation becomes satisfactory.

■ ***LipRead*'s Elliptical Classifier**

LipRead leverages 3 features to detect an attack: power in sub-50Hz, correlation coefficient, and amplitude skew. Analyzing the *False Acceptance Rate* (FAR) and *False Rejection Rate* (FRR), as a function of these 3 parameters, we have converged on a ellipsoidal-based separation technique. To determine the optimal decision boundary, we compute *False Acceptance Rate* (FAR) and *False Rejection Rate* (FRR) for each candidate ellipsoid. Our aim is to pick the parameters of the ellipse that minimize both FAR and FRR. Figure 13 plots the FAR and FRR as intersecting planes in a logarithmic scale (Note that we show only two features since it is not possible to visualize the 4D graph). The coordinate with minimum value along the canyon – indicating the *equal error rates* – gives the optimal selection of ellipsoid. Since it targets speech commands, this classifier is designed offline, one-time, and need not be trained for each device or individual.

5 Evaluation

We evaluate *LipRead* on 3 main metrics: (1) attack range, (2) inaudibility of the attack, and the recorded sound quality (i.e., whether the attacker's command sounds human-like), and (3) accuracy of the defense under various environments. We summarize our findings below.

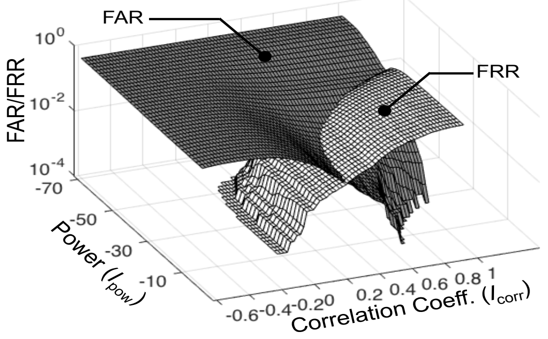


Figure 13: The False Acceptance Rate plane (dark color) and the False Rejection Rate plane (light color) for different sub-50Hz power and correlation values.

- We test our attack prototype with 984 commands to Amazon Echo and 200 commands to smartphones – the attacks are launched from various distances with 130 different background noises. Figure 15 shows attack success at 24ft for Amazon Echo and 30ft for smartphones at a power of 6$watt$.
- We record 12 hours of microphone data – 5 hours of human voice commands and 7 hours of attack commands through ultrasound speakers. Figure 16(c) shows that attack words are recognized by VEDs with equal accuracy as legitimate human words. Figure 16(b) confirms that all attacks are inaudible, i.e., the leakage from our speaker array is 5-10dB below human hearing threshold.
- Figure 17(a) shows the precision and recall of our defense technique, as 98% and 99%, respectively, when the attacker does not manipulate the attack command. Importantly, precision and recall remain steady even under signal manipulation.

Before elaborating on these results, we first describe our evaluation platforms and methodology.

5.1 Platform and Methodology

(1) Attack speakers: Figure 14(b) shows our custom-designed speaker system consisting of 61 ultrasonic piezoelectric speakers arranged as a hexagonal planar array. The elements of the array are internally connected in two separate clusters. A dual channel waveform generator (*Keysight 33500b series* [4]) drives the first cluster with the voice signal, modulated at the center frequency of 40kHz. This cluster forms smaller sub-clusters to transmit separate segments of the spliced spectrum. The second cluster transmits the pure 40kHz tone through each speaker. The signals are amplified to 30 Volts using a custom-made NE5534AP op-amp based amplifier circuit. This prototype is limited to a maximum power of 6$watt$ because of the power ratings of the operational amplifiers. More powerful amplifiers are certainly available to a resourceful attacker.

(2) Target VEDs: We test our attack on 3 different VEDs – Amazon Echo, Samsung $S6$ smartphone running Android v7.0, and Siri on an iPhone 5S running iOS v10.3. Unlike Echo, Samsung S-voice and Siri requires personalization of the wake-word with user's voice – adding a layer of security through voice authentication. However, voice synthesis is known to be possible [46, 5], and we assume that the synthesized wake-word is already available to the attacker.

Experiment setup: We run our experiments in a lab space occupied by 5 members and also in an open corridor. We place the VEDs and the ultrasonic speaker at various distances ranging up to 30ft. During each attack, we play varying degrees of interfering signals from 6 speakers scattered across the area, emulating natural home/office noises. The attack signals were designed by first collecting real human voice commands from 10 different individuals; MATLAB is used to modulate them to ultrasound frequencies. For speech quality of the attack signals, we used the open-source *Sphinx4* speech processing tool [1].

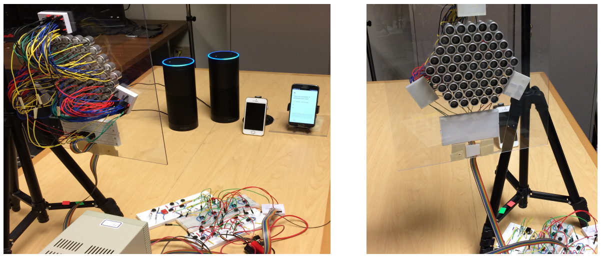

(a) (b)

Figure 14: *LipRead* evaluation setup: (a) Ultrasonic speaker and voice enabled devices. (b) The ultrasonic speaker array for attack.

5.2 Attack Performance

Activation distance: This experiment attempts to activate the VEDs from various distances. We repeatedly play the inaudible wake-word from the ultrasound speaker system at regular intervals and count the fraction of successful activation. Figure 15(a) shows the activation hit rate against increasing distance – higher hit-rates indicate success with less number of attempts. The average distance achieved for 50% hit rate is 24ft, while the maximum for Siri and Samsung S-voice are measured to be 27 and 30ft respectively.

Figure 15(b) plots the attack range again, but for the entire voice command. We declare "success" if the text to speech translation produces every single word in the command. The range degrades slightly due to the stronger need to decode every word correctly.

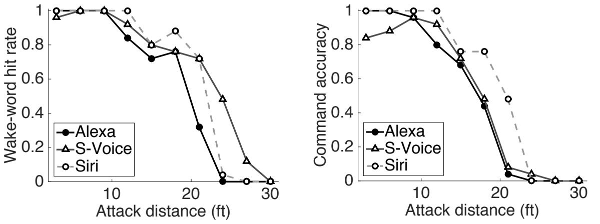
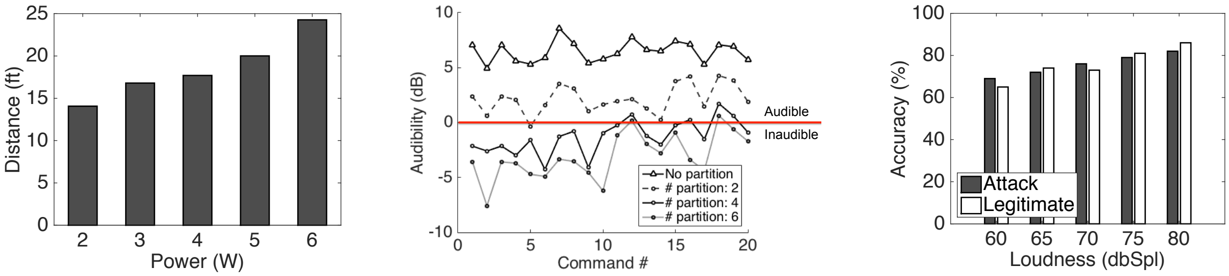


Figure 15: (a) The wake-word hit rate and (b) the command detection accuracy against increasing distances.

Figure 16(a) reports the attack range to Echo for increasing input power to the speaker system. As expected, the range continues to increase, limited by the power of our 6$Watt$ amplifiers. More powerful amplifiers would certainly enhance the attack range, however, for the purposes of prototyping, we designed our hardware in the lower power regime.

Leakage audibility: Figure 16(b) plots the efficacy of our spectrum splicing optimization, i.e., how effectively does *LipRead* achieve speaker-side inaudibility for different ultrasound commands. Observe that without splicing (i.e., "no partition"), the ultrasound voice signal is almost 5dB above the human hearing threshold. As the number of segments increase, audibility falls below the hearing curve. With 60 speakers in our array, we use 6 segments, each played through 5 speakers; the remaining 31 were used for the second $cos(2\pi f_c t)$ signal. Note that the graph plots the minimum gap between the hearing threshold and the audio playback, implying that this is a conservative worst case analysis. Finally, we show results from 20 example attack commands – the other commands are below the threshold.

Received speech quality: Given 6 speakers were transmitting each spliced segment of the voice command, we intend to understand if this distorts speech quality. Figure 16(c) plots the word recognition accuracy via Sphinx [1], an automatic speech recognition software. Evidently, *LipRead*'s attack quality is comparable to human quality, implying that our multi-speaker beamforming preserves the speech's structure. In other words, speech quality is not the bottleneck for attack range.



(a) (b) (c)

Figure 16: (a) Maximum activation distance for different input power. (b) Worst case audibility of the leakage sound after optimal spectrum partitioning. (c) Word recognition accuracy with automatic speech recognition software for attack and legitimate voices.

5.3 Defense Performance

Metrics: Our defense technique essentially attempts to classify the attack scenarios distinctly from the legitimate voice commands. We report the "*Recall*" and "*Precision*" of this classifier for various sound pressure levels (measured in *dBSPL*), varying degrees of ambient sounds as interference, and deliberate signal manipulation. Recall that our metrics refer to:

• *Precision:* What fraction of our detected attacks are correct?
• *Recall:* What fraction of the attacks did we detect?

We now present the graphs beginning with the basic classification performance.

Basic attack detection: Figure 17(a) shows the attack detection performance in normal home environment without significant interference. The average precision and recall of the system is 99% across various loudness of the received voice. This result indicates best case performance of our system with minimum false alarm.

Impact of ambient noise: In this section we test our defense system for common household sounds that can potentially mix with the received voice signal and change its features leading to misclassification. To this end, we played 130 noise sounds through multiple speakers while recording attack and legitimate voice signals with a smartphone. We replayed the noises at 4 different sound pressure levels starting from a typical value of 50 dBSPL to extremely loud 80 dBSPL, while the voice loudness is kept constant at 65 dBSpl. Figure 17(b) reports the precision and recall for this experiment. The recall remains close to 1 for all these noise levels, indicating that we do not miss attacks. However, at higher interference levels, the precision slightly degrades since the false detection rate increases a bit when noise levels are extremely high which is not common in practice.

Impact of injected noise: Next, we test the defense performance against deliberate attempts to eliminate nonlinearity features from the attack signal. Here the attackers strategy is to eliminate the $v^2(t)$ correlation by injecting noise in the attack signal. We considered four different categories of noise – white Gaussian noise to raise the noise floor, band-limited noise on the *Sub-50Hz* region, water-filling noise power at low frequencies to mask the correlated power variations, and intermittent frequencies below 50 Hz. As shown, in Figure 17(c), the process does not significantly impact the performance because of the power-correlation trade-off exploited by the defense classifier. Figure 17(d) shows that the overall accuracy of the system is also above 99% across all experiments.

6 Points of Discussion

We discuss several dimensions of improvement.

■ **Lack of formal guarantee:** We have not formally proved our defense. Although *LipRead* is systematic and transparent (i.e., we understand why it should succeed) it still leaves the possibility that an attack may breach the defense. Our attempts to mathematically model the self-convolution and correlation did not succeed since frequency and phase responses for general voice commands were difficult to model, as were real-world noises. A deeper treatment is necessary, perhaps with help from speech experts who can model the phase variabilities in speech. We leave this to future work.

■ **Generalizing to any signal:** Our defense is designed for the class of voice signals, which applies well to inaudible voice attacks. A better defense should find the true trace of non-linearity, not just for the special case of voice. This remains an open problem.

■ **Is air non-linear as well?** There is literature that claims air is also a non-linear medium [17, 10, 45]. When excited by adequately powerful ultrasound signals, self-convolution occurs, ultimately making sounds audible. Authors in [36, 2] are designing *acoustic spotlighting systems* where the idea is to make ultrasound signals audible only along a direction. We have encountered traces of air non-linearity, although in rare occasions. This certainly call for a separate treatment in the future.

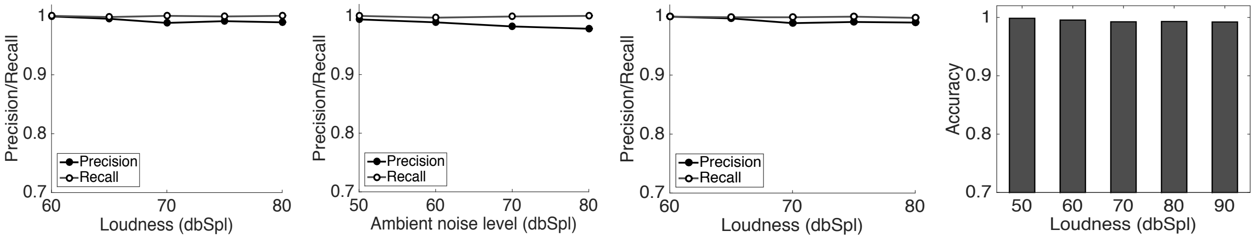


Figure 17: Precision and Recall of defense performance under various condition: (a) basic performance without external interference, (b) performance under ambient noise, and (c) performance under injected noise. (d) Overall accuracy across all experiments.

■ **Through-wall attack:** Due to the limited maximum power (6*watt*) of our amplifiers, we tested our system in non-blocking scenarios. If the target device is partially blocked (e.g. furnitures in the room blocking line-of-sight), the SNR reduces and our attack range will reduce. This level of power has not allowed us to launch through-wall attacks yet. We leave this to future work.

7 Related Work

■ **Attack on Voice Recognition Systems:** Recent research [11, 42] shows that spoken words can be mangled such that they are unrecognizable to humans, yet decodable by voice recognition (VR) systems. *GVS-Attack* [14] exploits this by creating a smartphone app that gives adversarial commands to its voice assistant. More recently, *BackDoor* [37] has taken advantage of the microphone's nonlinearity to design ultrasonic sounds which are inaudible to humans, but becomes recordable inside the off-the-shelf microphones. The application includes preventing acoustic eavesdropping with inaudible jamming signals. As follow up, [48, 39] show that the principles of *BackDoor* can be used to send inaudible attack commands to a VED, but requires physical proximity to remain audible. *LipRead* demonstrates the feasibility to increase the inaudible attack range, but more importantly, designs a defense against the inaudible attacks.

In past, researchers use near-ultrasound [27, 32, 40, 9, 21, 30] and exploited aliasing to record inaudible sound with microphone. A number of papers use other sound to camouflage audible signal in order to make it indistinguishable to human [24, 20, 12]. *CovertBand* [33] use music to hide audible harmonic components at the speaker. *LipRead*, on the other hand, use high frequency ultrasound as inaudible signal and leverages hardware nonlinearity to make them recordable to microphone.

■ **Acoustic Non-linearity:** A body of research [17, 10, 45], inspired by Westervelt's seminal theory [44, 43] on nonlinear Acoustics, studies the distortions of sound while moving through nonlinear mediums including the air. This raises the possibility that ultrasonic sound can naturally self-demodulate in the air to generate audible sounds, making it possible to develop a highly directional speaker [17, 10, 45]. Recently, *AudioSpotlight* [2], *SoundLazer* [7, 6], and other projects [47, 8, 34] have rolled out commercial products based on this concept. Ultrasonic hearing aids [29, 13, 15, 35, 31] and headphones [25] explore the human body as a nonlinear medium to enable voice transfer through bone conduction. Our work, however, is opposite of these efforts – we attempt to retain the inaudible nature of ultrasound while making it recordable inside electronic circuits.

■ **Speaker Linearization:** A number of research [23, 26, 18] studies the possibility of adaptive linearization of general speakers. Through simulations, the authors have shown that by pre-processing the input signal, they can achieve as much as 27*dB* reduction [18] of the nonlinear distortion in the noise-free case. Their techniques are not yet readily applicable to real speakers, since they have all assumed very weak nonlinearities, and over-simplified electrical and mechanical structures of speakers. With real speakers, especially ultrasonic piezoelectric speakers, it is difficult to fully characterize the parameters of the nonlinear model. Of course, if future techniques can fully characterize such models, our system can be made to achieve longer range with fewer speakers.

8 Conclusion

This paper builds on existing work to show that inaudible voice commands are viable from distances of 25+ ft. Of course, careful design is necessary to ensure the attack is truly inaudible – small leakages from the attacker's speakers can raise suspicion, defeating the attack. This paper also develops a defense against inaudible voice commands that exploit microphone nonlinearity. We show that non-linearity leaves traces in the recorded voice signal, that are difficult to erase even with deliberate signal manipulation. Our future work is aimed at solving the broader class of non-linearity attacks for any signals, not just voice.

Acknowledgement

We sincerely thank our shepherd Prof. Shyamnath Gollakota and the anonymous reviewers for their valuable feedback. We are grateful to the Joan and Lalit Bahl Fellowship, Qualcomm, IBM, and NSF (award number: 1619313) for partially funding this research.

References

[1] Cmu sphinx. `http://cmusphinx.sourceforge.net`. Last accessed 6 December 2015.

[2] Holosonics webpage. `https://holosonics.com`. Last accessed 28 November 2016.

[3] Inaudible voice commands demo. `https://www.youtube.com/watch?v=wF-DuVkQNQQ&feature=youtu.be`. Last accessed 24 September 2017.

[4] Keysight waveform generator. `http://literature.cdn.keysight.com/litweb/pdf/5991-0692EN.pdf`. Last accessed 24 September 2017.

[5] Lyrebird. `https://lyrebird.ai`. Last accessed 24 September 2017.

[6] Soundlazer kickstarter. `https://www.kickstarter.com/projects/richardhaberkern/soundlazer`. Last accessed 28 November 2016.

[7] Soundlazer webpage. `http://www.soundlazer.com`. Last accessed 28 November 2016.

[8] Woody norris ted talk. `https://www.ted.com/speakers/woody_norris`. Last accessed 28 November 2016.

[9] AUMI, M. T. I., GUPTA, S., GOEL, M., LARSON, E., AND PATEL, S. Doplink: Using the doppler effect for multi-device interaction. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing* (2013), ACM, pp. 583–586.

[10] BJØRNØ, L. Parametric acoustic arrays. In *Aspects of Signal Processing*. Springer, 1977, pp. 33–59.

[11] CARLINI, N., MISHRA, P., VAIDYA, T., ZHANG, Y., SHERR, M., SHIELDS, C., WAGNER, D., AND ZHOU, W. Hidden voice commands. In *USENIX Security Symposium* (2016), pp. 513–530.

[12] CONSTANDACHE, I., AGARWAL, S., TASHEV, I., AND CHOUDHURY, R. R. Daredevil: indoor location using sound. *ACM SIGMOBILE Mobile Computing and Communications Review 18*, 2 (2014), 9–19.

[13] DEATHERAGE, B. H., JEFFRESS, L. A., AND BLODGETT, H. C. A note on the audibility of intense ultrasonic sound. *The Journal of the Acoustical Society of America 26*, 4 (1954), 582–582.

[14] DIAO, W., LIU, X., ZHOU, Z., AND ZHANG, K. Your voice assistant is mine: How to abuse speakers to steal information and control your phone. In *Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices* (2014), ACM, pp. 63–74.

[15] DOBIE, R. A., AND WIEDERHOLD, M. L. Ultrasonic hearing. *Science 255*, 5051 (1992), 1584–1585.

[16] DOBRUCKI, A. Nonlinear distortions in electroacoustic devices. *Archives of Acoustics 36*, 2 (2011), 437–460.

[17] FOX, C., AND AKERVOLD, O. Parametric acoustic arrays. *The Journal of the Acoustical Society of America 53*, 1 (1973), 382–382.

[18] GAO, F. X., AND SNELGROVE, W. M. Adaptive linearization of a loudspeaker. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on* (1991), IEEE, pp. 3589–3592.

[19] GONZÁLEZ, G. G. G., AND NÄSSI, I. M. S. V. Measurements for modelling of wideband nonlinear power amplifiers for wireless communications. *Department of Electrical and Communications Engineering, Helsinki University of Technology* (2004).

[20] GRUHL, D., LU, A., AND BENDER, W. Echo hiding. In *International Workshop on Information Hiding* (1996), Springer, pp. 295–315.

[21] GUPTA, S., MORRIS, D., PATEL, S., AND TAN, D. Soundwave: using the doppler effect to sense gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), ACM, pp. 1911–1914.

[22] HAWKSFORD, M. J. Distortion correction in audio power amplifiers. *Journal of the Audio Engineering Society 29*, 1/2 (1981), 27–30.

[23] JAKOBSSON, D., AND LARSSON, M. Modelling and compensation of nonlinear loudspeaker.

[24] JAYARAM, P., RANGANATHA, H., AND ANUPAMA, H. Information hiding using audio steganography–a survey. *The International Journal of Multimedia & Its Applications (IJMA) Vol 3* (2011), 86–96.

[25] KIM, S., HWANG, J., KANG, T., KANG, S., AND SOHN, S. Generation of audible sound with ultrasonic signals through the human body. In *Consumer Electronics (ISCE), 2012 IEEE 16th International Symposium on* (2012), IEEE, pp. 1–3.

[26] KLIPPEL, W. J. Active reduction of nonlinear loudspeaker distortion. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings* (1999), vol. 1999, Institute of Noise Control Engineering, pp. 1135–1146.

[27] LAZIK, P., AND ROWE, A. Indoor pseudo-ranging of mobile devices using ultrasonic chirps. In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems* (2012), ACM, pp. 99–112.

[28] LEE, K.-L., AND MAYER, R. Low-distortion switched-capacitor filter design techniques. *IEEE Journal of Solid-State Circuits 20*, 6 (1985), 1103–1113.

[29] LENHARDT, M. L., SKELLETT, R., WANG, P., AND CLARKE, A. M. Human ultrasonic speech perception. *Science 253*, 5015 (1991), 82–85.

[30] LIN, Q., YANG, L., AND LIU, Y. Tagscreen: Synchronizing social televisions through hidden sound markers. In *INFOCOM 2017-IEEE Conference on Computer Communications, IEEE* (2017), IEEE, pp. 1–9.

[31] NAKAGAWA, S., OKAMOTO, Y., AND FUJISAKA, Y.-I. Development of a bone-conducted ultrasonic hearing aid for the profoundly sensorineural deaf. *Transactions of Japanese Society for Medical and Biological Engineering 44*, 1 (2006), 184–189.

[32] NANDAKUMAR, R., GOLLAKOTA, S., AND WATSON, N. Contactless sleep apnea detection on smartphones. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services* (2015), ACM, pp. 45–57.

[33] NANDAKUMAR, R., TAKAKUWA, A., KOHNO, T., AND GOLLAKOTA, S. Covertband: Activity information leakage using music. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1*, 3 (2017), 87.

[34] NORRIS, E. Parametric transducer and related methods, May 6 2014. US Patent 8,718,297.

[35] OKAMOTO, Y., NAKAGAWA, S., FUJIMOTO, K., AND TONOIKE, M. Intelligibility of bone-conducted ultrasonic speech. *Hearing research 208*, 1 (2005), 107–113.

[36] POMPEI, F. J. *Sound from ultrasound: The parametric array as an audible sound source*. PhD thesis, Massachusetts Institute of Technology, 2002.

[37] ROY, N., HASSANIEH, H., AND CHOUDHURY, R. R. Backdoor: Making microphones hear inaudible sounds. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services* (2017), ACM.

[38] SELF, D. *Audio power amplifier design handbook.* Taylor & Francis, 2006.

[39] SONG, L., AND MITTAL, P. Inaudible voice commands, 2017.

[40] SUN, Z., PUROHIT, A., BOSE, R., AND ZHANG, P. Spartacus: spatially-aware interaction for mobile devices through energy-efficient audio sensing. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services* (2013), ACM, pp. 263–276.

[41] TITZE, I. R., AND MARTIN, D. W. Principles of voice production. *The Journal of the Acoustical Society of America 104*, 3 (1998), 1148–1148.

[42] VAIDYA, T., ZHANG, Y., SHERR, M., AND SHIELDS, C. Cocaine noodles: Exploiting the gap between human and machine speech recognition. In *9th USENIX Workshop on Offensive Technologies (WOOT 15)* (Washington, D.C., 2015), USENIX Association.

[43] WESTERVELT, P. J. The theory of steady forces caused by sound waves. *The Journal of the Acoustical Society of America 23*, 3 (1951), 312–315.

[44] WESTERVELT, P. J. Scattering of sound by sound. *The Journal of the Acoustical Society of America 29*, 2 (1957), 199–203.

[45] YANG, J., TAN, K.-S., GAN, W.-S., ER, M.-H., AND YAN, Y.-H. Beamwidth control in parametric acoustic array. *Japanese Journal of Applied Physics 44*, 9R (2005), 6817.

[46] YE, H., AND YOUNG, S. High quality voice morphing. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on* (2004), vol. 1, IEEE, pp. I–9.

[47] YONEYAMA, M., FUJIMOTO, J.-I., KAWAMO, Y., AND SASABE, S. The audio spotlight: An application of nonlinear interaction of sound waves to a new type of loudspeaker design. *The Journal of the Acoustical Society of America 73*, 5 (1983), 1532–1536.

[48] ZHANG, G., YAN, C., JI, X., ZHANG, T., ZHANG, T., AND XU, W. Dolphinatack: Inaudible voice commands. *arXiv preprint arXiv:1708.09537* (2017).

[49] ZHANG, G., YAN, C., JI, X., ZHANG, T., ZHANG, T., AND XU, W. Dolphinattack: Inaudible voice commands. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)* (2017), ACM.