Patronus: Preventing Unauthorized Speech Recordings with Support for Selective Unscrambling

Lingkun Li[1*], Manni Liu[1*], Yuguang Yao[1], Fan Dang[2], Zhichao Cao[1], Yunhao Liu[1,2]
[1]Michigan State University [2]Tsinghua University
{lilingk1,liumanni,yaoyugua,caozc}@msu.edu,{dangf09,yunhaoliu}@gmail.com

ABSTRACT

The widespread adoption and ubiquity of smart devices equipped with microphones (*e.g.,* cellphones, smartwatches, *etc.*) unfortunately create many significant privacy risks. In recent years, there have been several cases of people's conversations being secretly recorded, sometimes initiated by the device itself. Although some manufacturers are trying to protect users' privacy, to the best of our knowledge, there is not any effective technical solution available. In this work, we present Patronus, a system that can both prevent unauthorized devices from making secret recordings while allowing authorized devices to record conversations. Patronus prevents unauthorized speech recording by emitting what we call a *scramble*, a low-frequency noise generated by inaudible ultrasonic waves. The scramble prevents unauthorized recordings by leveraging the nonlinear effects of commercial off-the-shelf microphones. The frequency components of the scramble are randomly determined and connected with linear chirps, and the frequency period is fine-tuned so that the scramble pattern is hard to attack. Patronus allows authorized speech recording by secretly delivering the scramble pattern to authorized devices, which can use an adaptive filter to cancel out the scramble. We implement a prototype system and conduct comprehensive experiments. Our results show that only 19.7% of words protected by Patronus' scramble can be recognized by unauthorized devices. Furthermore, authorized recordings have 1.6x higher perceptual evaluation of speech quality (PESQ) score and, on average, 50% lower speech recognition error rates than unauthorized recordings.

CCS CONCEPTS

• **Networks** → *Mobile networks*; • **Computer systems organization** → **Special purpose systems**; • **Security and privacy** → **Privacy-preserving protocols**; • **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**.

KEYWORDS

Privacy Protection; Microphone; Nonlinear Effects

[*] Co-primary authors.

SenSys '20, November 16–19, 2020, Virtual Event, Japan

ACM ISBN 978-1-4503-7590-0/20/11…$15.00
https://doi.org/10.1145/3384419.3430713

ACM Reference Format:
Lingkun Li, Manni Liu, Yuguang Yao, Fan Dang, Zhichao Cao, and Yunhao Liu. 2020. Patronus: Preventing Unauthorized Speech Recordings with Support for Selective Unscrambling. In *The 18th ACM Conference on Embedded Networked Sensor Systems (SenSys '20), November 16–19, 2020, Virtual Event, Japan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3384419.3430713

1 INTRODUCTION

Human beings have long used acoustic signals to exchange information with each other. Human beings now use acoustic signals, which is speech, to exchange information with ubiquitous smart devices such as smartphones, smartwatches, and digital assistants that are equipped with embedded microphones. While these speech detection and recognition capabilities make possible many convenient features, they also introduce many privacy risks such as secret, unauthorized recordings of our private speech [1, 2] that can have real world consequences. For example, the Ukrainian prime minister offered his resignation after an unauthorized recording was leaked [3].

Manufacturers claim that they are trying their best to protect users' privacy, but there is no effective and user-friendly technical anti-recording solution available despite the fact that anti-recording is not a new problem. One existing anti-recording solution is to talk near a white noise source, *e.g.,* near an FM radio tuned to unused frequencies, so that the conversation cannot be clearly recorded. This approach is not user-friendly because the people having the conversation must put up with the white noise that interferes with their normal communication. A similar solution [4] emits high frequency noise near the upper bound of human sensitivity; most people do not notice the interference, but pets and infants may notice it [5], so this solution is not environment-friendly. Electromagnetic interference was an effective anti-recording solution [6] in the past, but modern microphones are immune to electromagnetic interference. Moreover, all of these traditional anti-recording approaches cannot allow authorized devices to clearly record conversations.

Any effective anti-recording solution must provide the following three key properties: (1) normal human conversation should be unaffected by the anti-recording solution meaning the anti-recording solution should not change what humans hear while having a conversation; (2) unauthorized devices should not be able to make a clear recording of any conversation protected by the anti-recording solution; (3) authorized devices should be able to make a clear recording of any conversation protected by the anti-recording solution.

One potential solution that can satisfy all three properties is to generate multiple ultrasonic frequency sound waves because of the

following two properties of ultrasonic waves. First, humans cannot hear ultrasonic sound waves. Second, commercial off-the-shelf (COTS) microphones exhibit nonlinear effects, which means that when these microphones receive multiple ultrasonic sound waves, they generate low-frequency sound waves that can be heard by humans and thus interfere with the clarity of recordings made with those microphones [5, 7–12]. There are three main challenges that must be overcome in order to develop an ultrasonic anti-recording solutions that satisfies the three key properties:

(1) First, any ultrasonic anti-recording solution must defend against potential attacks such as using Short-time Fourier transform (STFT) to analyze unauthorized recordings and using filters to cancel out the low-frequency sound waves that interfere with recording clarity.
(2) Second, ultrasound travels along a straight line [13], which means a single ultrasonic wave generator can only interfere with recording devices within a limited range of angles from the generator. In practice, it is difficult to design an ultrasonic anti-recording solution that can neutralize all recording devices within a large coverage area.
(3) Finally, the performance of authorized devices could be affected by the ringing effect due to electronic behaviors. Such ringing impulses are hard to be canceled and may remain in authorized recordings, severely downgrading the quality of the descrambled recordings.

In this paper, we present Patronus, an ultrasonic anti-recording system that satisfies the three key properties. Patronus has two key components: the *scramble* that is the pseudo-noise generated at all microphones, and *descrambling* that is the process to remove the scramble for authorized devices. We form the scramble by randomly picking frequencies from the human voice frequency band and then shifting them to the ultrasonic band. To thwart STFT attacks, we further fine-tune the period of the scramble so that it cannot be easily analyzed and canceled. We add a reflection layer with a curved surface to create a reflected ultrasonic wave that can cover a wider area. Finally, to mitigate ringing effects, *i.e.,* sudden hardware impulses due to discrete frequency changes of current waves, we use chirps to smooth the frequency changing components of the scramble, as shown in Figure 1.

Patronus lets authorized devices clearly record audio conversations by sending them the scramble pattern. With scramble pattern, the authorized device applies the Normalized Least-Mean-Square (NLMS) adaptive filter [14] to cancel the scramble and thus produce a clear audio recording of the conversation.

We implement a prototype of Patronus and conduct comprehensive experiments to evaluate its performance. We use the Perceptual Evaluation of Speech Quality (PESQ) [15], the Speech Recognition Vocabulary Accuracy (SRVA, see Section 6), and speech recognition error rates (1 - SRVA) to evaluate the performance of Patronus. Our results show that only 19.7% of the words protected by Patronus' scramble can be recognized by unauthorized devices. Furthermore, authorized recordings have 1.6x higher PESQ and, on average, 50% lower speech recognition error rates than unauthorized recordings.

In this paper, we provide several unique technical contributions when compared to existing works. First, to the best of our knowledge, Patronus is the first system to leverage the nonlinear effect of COTS microphones to prevent unauthorized recordings while allowing authorized recordings. Second, we perform a thorough study of the nonlinear effects of ultrasound frequencies including the effects of higher orders whereas recent works[7–9, 16] only consider the order up to 2. This is critical for descrambling when the signal components with order higher than 2 will likely lie in the human voice frequency band, which means simply cutting off the high frequency components will result in message loss. Instead, our descrambling solution carefully removes these higher order frequencies using an NLMS filter. Third, we mitigate ringing effects by connecting scramble segments with chirps. This simplifies learning the coefficients of impulse response in existing work [7], especially when we deploy multiple ultrasonic transducers in a large space. In general, our contributions are as follows:

- We propose a novel ultrasound modulation approach to provide privacy protection against unauthorized recordings that does not disturb normal conversation.
- We do a thorough study around the nonlinear effect of ultrasound on commercial microphones and propose an optimized configuration to generate the scramble.
- To overcome the fact that ultrasound travels in a straight line, we design a low cost reflection layer to effectively enlarge the coverage area of Patronus in a cost-effective way.
- We present Speech Recognition Vocabulary Accuracy, a new metric to measure the recording quality. Our experimental results with both PESQ and SRVA show that Patronus effectively prevents unauthorized devices from making secret recordings.

The organization of the rest of this paper is as follows. Section 2 introduces related work. Section 3 introduces the nonlinear effect of common microphones, which we analyze more thoroughly than existing works. Section 4 presents the design of Patronus. Section 5 presents the prototype implementation of Patronus. Section 6 presents our evaluation results of Patronus. Section 7 discusses the limitations of Patronus and future work, and Section 8 concludes this work.

2 RELATED WORKS

2.1 Nonlinear Effect of Microphones

There has been a lot of research into the nonlinear effect of microphones. For many years, the development of ultrasonic systems on smartphones was restricted due to being limited to a roughly 4 kHz range of frequencies between the high end of human hearing to the cutoff frequency of typical microphones. Furthermore, some infants and pets can actually perceive frequencies within this small band. Roy *et al.* [7] performed detailed research on the nonlinear effects of microphones to break through these limitations and expand the working frequency band for ultrasonic systems on smartphones. DolphinAttack [9] leverages the nonlinear effect to generate audio commands that are inaudible to humans. After being recorded by the microphone, the input ultrasonic signals would generate a shadow signal that could be recognized by VCS. Therefore, attackers can perform unauthorized commands without being discovered. SurfingAttack [12] uses oscillation of a surface such as a table to transmit inaudible commands. With this modality, attackers can deploy their speakers in hidden spots such as the back of the

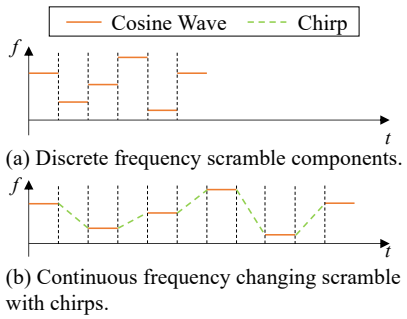


Figure 1: Using chirps to smooth the frequency changing components of the scramble.

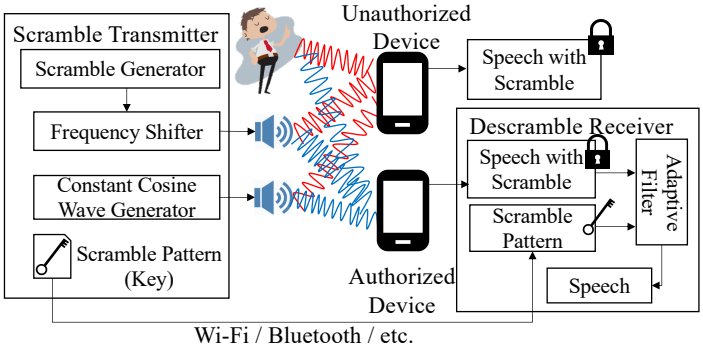


Figure 2: System Overview.

surface being used to transmit the secret commands. LipRead [8] extends the attack range by leveraging characteristics of human hearing. It also puts forward a model to filter out such commands generated by the nonlinear effect. Metamorph [10] injects inaudible commands into human-made commands to achieve unauthorized actions. AIC [16] presents a mechanism that fundamentally cancels inaudible commands against VCS, which we will discuss as an attack model in Section 4.2. NAuth [11] uses the nonlinear effect to authenticate devices. Unlike most of these methods, Patronus aims to preserve privacy by adding a removable scramble generated by ultrasonic signals to the recorded human speech. From a technical perspective, Patronus is unique in that it takes into account third and higher order terms from the nonlinear effect. Our experiments show those high order terms can affect recordings whereas most existing methods (e.g., AIC) only consider the second order term and assume the higher order sub-band of the microphone is clean.

2.2 Dual Channel Applications

Some applications leverage the difference between humans and devices. For example, human eyes and devices have different perceptions of flicker frequency. Technologies exist that use this phenomena to communicate between the screen and the camera without affecting human vision [17–20]. Likewise, some technologies modulate acoustic signals in ways that no human can detect to communicate between devices [21, 22].

The difference between the sensitivity of humans and devices is also used in privacy protection. Kaleido [23] protects a movie's copyright by adding a flashing distractor with very high frequency into movie frames that cannot be seen by human eyes. If such a protected movie is subsequently recorded by an unauthorized camera equipped with a rolling shutter, the distractor will be visible on the unauthorized recording because of its high sample rates making the pirated recording a low quality recording. LiShield [24] also uses the Rolling Shutter effect to reduce the quality of photos. Lights with different colors are set to flash in alternating high frequencies that provide normal lighting because human eyes cannot sense the flashing. However, cameras are influenced because the Rolling Shutter samples column by column meaning unexpected color stripes will appear on the photo. In the end, it prevents unauthorized cameras from taking photos. Although Patronus has a similar motivation to prevent unauthorized recordings, Patronus is different from the two papers as it targets acoustics rather than visuals.

3 NONLINEAR BEHAVIOR OF COMMON MICROPHONES

In this section, we provide a brief primer about nonlinearity of common microphones; a more comprehensive introduction can be found in recent papers [7, 8]. Ideally, COTS microphones are linear systems. Given the input signal $s(t)$, the output signal $y(t)$ is expected to be linear combinations of the input signal, i.e., $y(t) = A_1 s(t)$ where A_1 is the complex gain quantifying the change of the phase and amplitude. Due to the physical properties of materials and variations in manufacturing, the components of a common microphone, such as the diaphragm and the pre-amplifier, are imperfect and typically do not constitute a linear system. As a result, COTS microphones, which are widely equipped on smartphones and smartwatches, typically exhibit nonlinear behavior. Specifically, the output signal $y(t)$ is under nonlinear effect, where $y(t) = A_1 s(t) + A_2 s^2(t) + A_3 s^3(t) + \cdots$, and the power gains of each component satisfy $|A_m| > |A_n| (m < n)$.

When the input signals are composed of two different ultrasonic frequencies, the output from a nonlinear microphone would contain several new shadow sounds with frequencies that are a linear combination of the two input frequencies. Assuming that the input signal is $s(t) = \cos(2\pi f_1 t) + \cos(2\pi f_2 t)$ where f_1 and f_2 are the ultrasonic frequencies, the output signal would be $y(t) = \sum_{i=1}^{+\infty} A_i s^i(t)$. Without loss of generality, we assume $f_1 > f_2$ in the following discussion. For each component $A_i s^i(t)$,

$$
\begin{aligned}
s^i(t) &= (\cos(2\pi f_1 t) + \cos(2\pi f_2 t))^i \\
&= \mu + \sum_{j=1}^{i} [\alpha_j \cos(2\pi j f_1 t) + \beta_j \cos(2\pi j f_2 t)] \\
&+ \sum_{j=1}^{i-1} [\lambda_j \cos(2\pi (j f_1 - (i-j) f_2) t) + \gamma_j cos(2\pi (j f_1 + (i-j) f_2) t)],
\end{aligned}
$$

where α_j, β_j, λ_j and γ are coefficients of the polynomial expansion, and μ is the consequent constant.

After the pre-amplifer, the signals would pass through an embedded low-pass filter whose cut-off frequency is usually 24 kHz. Since

f_1 and f_2 are both ultrasonic frequencies, jf_1 and jf_2 are all ultrasonic frequencies. However, if $i = 2j$, $jf_1 - (i-j)f_2 = j(f_1 - f_2)$ may be a non-ultrasonic frequency when j is small enough. Therefore, when the input signal is $s(t) = \cos(2\pi f_1 t) + \cos(2\pi f_2 t)$, new audible cosine waves $\cos(2\pi j(f_1 - f_2)t)$ appear, where $j = 1, 2, \ldots, k$, $k \leq i$, and $k(f_1 - f_2) \leq 24\,\text{kHz}$. Existing works like BackDoor[7] and DolphinAttack[9] make use of $A_2 s^2(t)$ but ignore higher-order components; they essentially assume that for $i > 2$, $|A_i|$ is relatively small and has little effect on the output signal. However, in our experiments, we find that more high-order components should be taken into consideration as they do affect the output signal.

4 DESIGN

4.1 Overview

As shown in Figure 2, there are three parties involved in Patronus: the Scramble Transmitter, authorized devices with descramble receivers, and unauthorized devices.

The Scramble Transmitter sends a series of scramble signals with randomly varying frequencies. To ensure that unauthorized voice recordings will be affected, the frequencies of the recorded scrambles should be located in the human voice band. Therefore, we use the Scramble Generator to generate random frequencies in the target range, store them as a secret key, and send them to the Descramble Receivers through Wi-Fi, Bluetooth, or other media. The Scramble Generator then generates cosine wave segments according to these frequencies. The generated segments are then sent to the Frequency Shifter and their frequencies will be increased by f_0, which is an ultrasonic frequency. To ensure the scramble signal is picked up by microphones of unauthorized devices because of the nonlinear effect, we design a Constant Cosine Wave Generator to transmit a cosine wave with a constant ultrasonic frequency of f_0.

During human talking protected by Patronus, the actual human conversation plus two ultrasonic signals will arrive essentially simultaneously at recorders (both authorized and unauthorized) and human ears. Human ears will not detect the ultrasonic signals and thus receive the human conversation with no additional noise. As discussed in Section 3, the two ultrasonic signals will generate a shadow audible signal that will be included in any recording made by a COTS microphone due to nonlinear effects. This applies to both authorized and unauthorized devices. Authorized devices, which receive a secret key from the Scrambling Transmitter, can generate the scramble waveform. They can then feed the scramble waveform along with the scrambled recording into an adaptive filter to extract clear speech from the scrambled speech. The details of descrambling will be discussed in Section 4.5.

We must overcome three challenges in order to design Patronus. First, we must design a system whose working area is as large as possible. This is difficult because a sound wave of high frequency typically travels along a straight line meaning a straightforward implementation of ultrasonic generators will only cover a small area defined by a limited range of angles. Second, there is a trade-off between a shorter and a longer period of scramble frequencies. As the period increases, the system is more vulnerable to unauthorized recordings using STFT attacks. As the period decreases, the difficulty of descrambling increases. Our goal is to maximize the information recovered by authorized devices over unauthorized ones without exposing the scramble pattern to STFT. These details are discussed in Section 4.3.4. Third, when frequency changes frequently, a severe ringing effect (Section 4.3) occurs in the scrambled recording, which affects even the recordings made by authorized devices after descrambling. We use chirps to connect each frequency component of the scramble to eliminate the sudden change of the input to ultrasonic transducers, hence minimizing the ringing effect and enhancing the quality of the recovered speech by authorized devices.

4.2 Attack Model

Based on common acoustic processing technologies and known properties of nonlinearity effects, we consider the following types of attacks:

4.2.1 Short-Time Fourier Transform (STFT). One natural way for an unauthorized device to try to extract a useful recording from its scrambled recording is to analyze the scrambled recording with STFT and filter out suspicious frequencies. We address this attack model by changing the scramble frequency according to a finely-tuned period model, making it impossible for the attacker to obtain each exact scramble frequency along with its start and end time. Detailed analysis is provided in Section 4.3.4. Even with the correct scramble frequencies available, bandpass filters will not work because the scramble frequencies are selected from the human voice band. The frequencies from chirps and those from human speaking are mixed together. To prove Patronus can defeat this attack model, we simulate the attack scenario when (1) the attacker is aware that our scramble pattern is varying continuous waves smoothed by chirps (2) the attacker calculates approximate scramble frequencies with STFT (3) the attacker applies NLMS adaptive filter (Section 4.5.4) to remove the scramble with the approximate scramble frequencies they obtained from STFT. Our simulated attack experiments, provided in Section 6.8, show that this attack will fail because the approximate scramble frequencies are not accurate enough.

4.2.2 Extra Ultrasonic Transmitter Attack. After DolphinAttack[9] proposes to inject malicious commands into ultrasound, AIC [16] adds three more ultrasonic transmitters to cancel the malicious commands and protect Voice Control Systems (VCS). AIC assumes the legitimate as well as malicious commands are within the lower sub-band of the microphone sensible frequency band. Their added ultrasonic transmitters project only the malicious commands onto the higher sub-band, which can be used to filter the malicious commands in the low sub-band. With a fast changing of scramble frequencies, we can cover the whole frequency band, and make sure no clean band is left for attackers.

4.2.3 Wi-Fi/Bluetooth Snifing. Attackers can sniffer the Wi-Fi or Bluetooth channel to get the scramble pattern transmitted from the Scramble Transmitter to the authorized device. However, there are many cryptographic approaches to prevent attackers from sniffing channels. For example, we can encrypt the scramble pattern by AES-CTR using a pre-shared key and then directly send it to authorized devices.

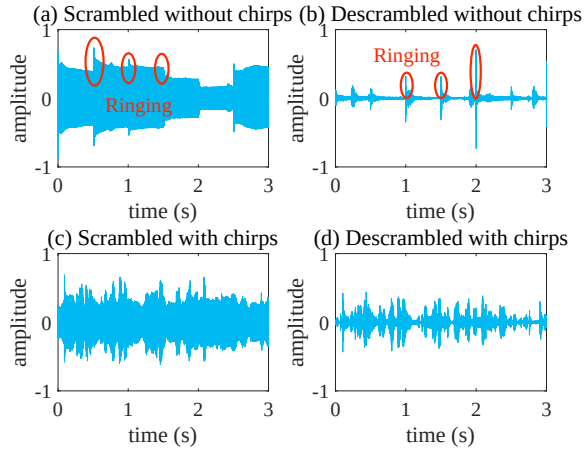


Figure 3: Illustration of how linear chirps mitigate the ringing effect.

4.2.4 Physical Attacking. There are also some physical attack models. First, attackers can place an obstacle before the Scramble Transmitter. However, attackers cannot do it secretly and nobody would like to do so. Second, attackers may just wrap a cover on their microphones. However, the cover itself may defeat the attackers objective of making a good recording. Although Patronus cannot perfectly handle such attack models, it enhances the difficulty of making an unauthorized recording. Finally, attackers may conduct experiments to discover where Patronus fails. This can be fixed by enlarging the working area through some methods that we will discuss later.

4.3 Ultrasonic Scramble Modulation

Two ultrasonic signals will be superimposed at the recorders to create the desired low-frequency component. In the design of the scramble using ultrasonic signals, we mainly consider the following issues:

4.3.1 Range of Frequency. The first issue is how to make it hard to cancel out the scramble without the key. Basically, the range of human speech frequency is from 85 Hz to 255 Hz [25, 26]. If the scramble consists of multiple random frequencies from this range, it is hard for attackers to cancel the scramble using linear filters. The application of a linear filter, *e.g.,* highpass filter, will not only cancel the scramble, it will also change the original human speech. To ensure the scramble covers all human speech frequencies in practice, we modulate the scramble with a wider frequency band than $[85, 255]$ Hz.

4.3.2 Random Frequencies. If we always use specific frequencies to generate the scramble, attackers could analyze the frequency spectrum of their recordings to infer the scramble frequencies; with those, they could then recover the original audio signals. To address this issue, we choose scramble frequencies randomly. We also periodically change the scramble frequencies over time. The sequence of scramble frequencies can be thought of as a one-time pad key. Without the sequence, it would be difficult for attackers to remove the scramble.

4.3.3 Ringing Effect. Frequent changing of the scramble frequencies produces a ringing effect [7] that makes it challenging for authorized devices to produce a high-quality descrambled recording. Specifically, the ringing effects incur heavy-tailed impulse responses that will remain in descrambled recordings as shown in Figure 3 (a) and (b). Since the ringing effect occurs when the input changes suddenly, we use a chirp signal to connect two adjacent segments with different frequencies in the scramble to smooth such a sudden change. Specifically, when the scramble changes from frequency A to frequency B, we add a transition signal that starts at frequency A and moves linearly to end with frequency B.

The impulse incurred by ringing effects can have a very high amplitude or power. It will suppress other signals due to the microphone Passive Gain Suppression [7]. Figure 3 confirms that the ringing effect is mitigated by chirps. Figure 3 (a) shows a scrambled recording with no chirp, the resulting descrambled recording in Figure 3 (b) has many areas where most of the signal is suppressed. In contrast, Figure 3 (c) exhibits a scrambled recording with chirp signals, the resulting descrambled recording in Figure 3 (d) does not have the peak signals corresponding to the ringing effect and the rest of the signal is not suppressed.

4.3.4 Duration of each frequency. The next challenge is choosing the proper duration for each frequency in the sequence of scramble frequencies. Intuitively, if we give each frequency a long duration, unauthorized devices could easily split the record into multiple segments where each segment is only protected by a constant frequency scramble. They could then apply simple techniques such as using a linear bandpass filter to the scrambled recording to extract a clear speech recording.

More generally, there are two competing issues in choosing the duration of each scramble frequency, namely, defending against STFT attacks that are discussed in Section 4.2.1, and ensuring that authorized devices can obtain high-quality descrambled recordings. We first consider defending against STFT attacks. An STFT attack can successfully remove the scramble waveform if it can both accurately infer the frequencies and time periods for each scramble frequency in the sequence of frequencies. When the window length is n, the frequency resolution would be $\Delta f = \frac{f_s}{n} = \frac{f_s}{f_s \times t} = \frac{1}{t}$ where f_s is the sampling rate and t is the duration of the window. Taking 0.1s as an example, the offset of STFT can reach 10Hz. If the attacker tries to improve the frequency resolution by lengthening the window, the accuracy of the estimated time periods for the given scramble frequency will diminish. If the scramble frequency duration is long, scramble frequency will exhibit fewer changes within any given window, thus STFT attacks can use longer windows to accurately estimate the frequency with exact estimates of the frequency time period. Therefore, to thwart STFT attacks, we should make the frequency duration as short as possible. However, a too-short duration may misshape the scrambled recording due to imperfect hardware. A typical microphone and speaker use a diaphragm to sense and generate the vibration; this diaphragm moves continuously and can not change its position instantaneously. Circuit latency also makes it hard for the system to respond to frequent and instant changes. As a result, the scrambled waveform would be slightly distorted. This means the NLMS adaptive filter at authorized

devices may not correctly descramble the scrambled waveform because it does not expect the distortion caused by frequent frequency changes. Therefore, the frequency duration cannot be too short. In summary, to balance these competing concerns, we must find a frequency duration that maximizes the information recovered by authorized devices compared to the information recovered by unauthorized devices. To identify a good frequency duration, we measure the descrambling performance with different frequency durations in Section 6.8.

4.3.5 Key Construction. We have two choices to construct the key for granting the privilege of recording the audio to authorized devices. One is directly using the scramble waveform generated by the Scramble Generator as the key. After getting the scramble waveform, authorized devices remove the scramble from the recorded audio. But there are some issues we need to consider. First, the sampling rate of authorized devices may vary from one to another. It means that in terms of the digital signal, devices having different sampling rates will get different presentations of the same scramble waveform. To grant the privilege to devices, the Scramble Transmitter should generate different digital scramble waveforms according to different sampling rates of authorized devices. This results in high computational overheads. Second, in addition to different sampling rates from different authorized devices, the sampling rates of the Scramble Generator and an authorized device may be also different. As a result, the scramble that the speaker emitted might have a different presentation of the recorded waveform.

In Patronus, we choose another way to construct the key. We select the frequency sequence used to generate the scramble as the key. After receiving the frequency sequence, an authorized device can reconstruct the scramble waveform with their sampling rates, which we discuss in more detail later. After that, an authorized device can use the reconstructed scramble waveform to remove the scramble from the recording and get the clear speech.

With the discussion above, we formally describe the scramble generation. We set one speaker to transmit an ultrasonic continuous wave $S_1(t) = \cos(2\pi f_0 t)$, while the other speaker transmits continuous waves linked by chirps $S_2(t) = \cos(2\pi f(t)t)$, where

$$f(t) = \begin{cases} f_i, & (2i-2)\Delta t \le t < (2i-1)\Delta t, \\ f_i + \frac{f_{i+1}-f_i}{\Delta t}t, & (2i-1)\Delta t \le t < 2i\Delta t, \end{cases} \quad (1)$$

and $f_i (i = 1, \ldots, n)$ are randomly generated constant frequencies. Δt is the duration of a single sine wave or a chirp. The induced low-frequency noise will be

$$R(t) = \cos(2\pi(f(t) - f_0)t). \quad (2)$$

To ensure $R(t)$ covers human voice, $f_i (i = 1, \ldots, n)$ are sampled from $[f_{low} + f_0, f_{high} + f_0]$ where $[f_{low}, f_{high}]$ covers the human voice band.

4.4 Enlarge Scramble Working Area

The scramble signal is generated by two ultrasonic signals, which incurs another issue as the ultrasonic wave typically propagates in a straight line. In other words, if you want to prevent a certain device from recording, the ultrasonic transducers should be pointed directly towards that device. This results in a limited coverage area for ultrasonic anti-recording solutions.

Inspired by lamps that often use a bow-shaped cover to reflect the light beam in many directions, we build a reflection layer that reflects the ultrasonic wave in many directions. As Figure 4 shows, we put ultrasonic transducers near the center of the reflection layer and place the devices (authorized and unauthorized) in the working area. When the ultrasonic wave hits the reflection layer, it gets reflected in many directions leading to a much larger cover area.

4.5 Grant Recording Privilege

The goal of Patronus is not only to block unauthorized devices from recording audio, but also to provide authorized devices with a mechanism to recover speech. Patronus achieves this by creating a way for authorized devices to remove the scramble from the scrambled recording. Specifically, Patronus grants the clear recording privilege to authorized devices using the following steps.

4.5.1 Key Transmission. The Descramble Receiver needs the waveform of the scramble generated by the Scramble Generator before it can remove the scramble. Intuitively, if it had the pure scramble waveform, it could remove the scramble from the recorded audio by subtracting the scramble waveform from the recorded audio waveform. The scramble waveform here acts as the key for deciphering the recorded audio. We send the key through non-acoustic channels such as Wi-Fi or Bluetooth with cryptographic protection to prevent eavesdroppers from getting the key. Additionally, because of the randomness of scramble frequencies, they cannot get a usable scramble waveform by listening to the acoustic channel. Instead, they can get either the combination of interfered speech with scramble, or get the scramble without speech but independent of the successive scramble waveform.

4.5.2 Scramble Reconstruction. As discussed in Section 4.3, the Scramble Transmitter sends the random frequency sequence instead of the scramble waveform to authorized devices as the key. Patronus needs to use these frequencies to reconstruct the scramble waveform before removing the scramble. An authorized device uses Equation (2) and its recording sampling rate to generate the scramble waveform.

4.5.3 Synchronization. We need to synchronize the reconstructed scramble with the recorded scramble before removing it from recordings. Specifically, we choose a segment from the reconstructed scramble as the template, *e.g.,* the beginning segment. Then we use cross-correlation to find the segment that is the most similar to the template. We then synchronize the recorded scramble and the reconstructed scramble by aligning the two segments.

4.5.4 Adaptive Filtering. Now we have the waveform of the scramble. The next task is to remove the scramble from the recorded audio with the known waveform of the scramble. Practically, we cannot directly subtract the scramble from the recorded audio because when the sound propagates through the air, it will be distorted due to reflection and attenuation. We use adaptive filter to remove the waveform-known scramble.

Adaptive filter is widely used in Active Noise Cancellation (ANC) headsets. Technically, there is a reference microphone outside the headset. The reference microphone captures the noise, and the digital signal processor (DSP) generates the anti-noise wave according

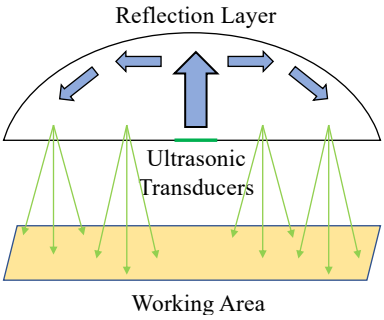


Figure 4: Enlarge working area with reflection.

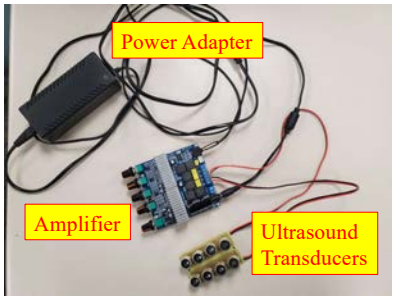


Figure 5: Implementation of Scramble Transmitter.

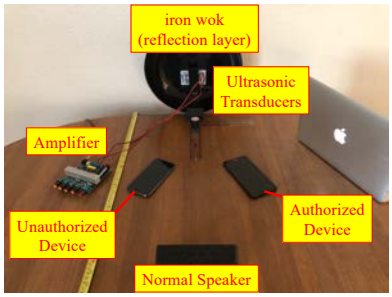


Figure 6: Prototype of Patronus

to the captured noise. When the noise wave and the anti-noise wave arrive at the ear, they eliminate each other. In Patronus, we denote the speech as x_1. It propagates through the acoustic channel h_1, arrives at the authorized device and becomes $h_1 * x_1$, where the operator $*$ denotes the convolution operation. Additionally, we denote the scramble waveform that is generated by non-linear effects and recorded by the authorized device as x_2. It propagates through another channel h_2, arrives at the authorized device and becomes $h_2 * x_2$. Therefore, the audio recorded by the authorized device is

$$y = h_1 * x_1 + h_2 * x_2. \tag{3}$$

Similar to ANC headsets, here we see the scramble x_2 as the noise in ANC headsets. Different from ANC headsets, the noise here is generated from the key as we discussed in Section 4.5.2. Therefore, we can use the Normalized Least-Mean-Square (NLMS) Adaptive Filter [14] to remove the scramble. Formally, we are trying to find a channel vector h_2' to solve the optimization problem

$$\min E[(y - h_2' * x_2)^2]. \tag{4}$$

When the expectation in Equation (4) is minimized, $h_2 \approx h_2'$. Therefore, $h_1 * x_1 \approx y - h_2' * x_2$, and it can be regarded as the speech without the scramble. Stochastic gradient descent is usually adopted to solve the optimization problem defined by Equation (4), but it is hard to derive the gradient of the expectation. Researchers thus use $(y - h_2' * x_2)^2$ instead of the expectation to solve the problem. In this way, the noise gets canceled [27].

Following this design, we can develop a mechanism that prevents unauthorized recording while supporting authorized recording. The mechanism also prevents attackers from descrambling without authorization. Figure 7 gives an example. A piece of VOA news audio is used as the original record, the attack result has severe scramble effects just like the unauthorized record, but the authorized record removes almost all scrambles.

5 IMPLEMENTATION

This section discusses the details of the implementation of Patronus, which contains two parts, the Scramble Transmitter and the Descramble Receiver for authorized devices. We use an ordinary smartphone with its built-in audio recorder as the Unauthorized Device or Authorized Device.

5.1 Scramble Transmitter

5.1.1 Hardware Implementation. As Figure 5 shows, we use eight TCT40-16R/T 16 mm ultrasonic transducers. Half of them play the frequency-shifted scramble and they are connected in parallel. The other half play the fixed-frequency cosine wave and are connected in parallel as well. We utilize an AOSHIKE DC12V-24V 2.1 Channel TPA3116 Subwoofer Amplifier Board to enhance the power of output ultrasonic signals. The two waveforms are played through a stereo channel. The frequency-shifted scramble uses the left channel, and the constant-frequency cosine wave uses the right channel.

As we have discussed in Section 4.4, we use a reflection layer to enlarge the working area. In this prototype, we use an iron wok as the reflection layer. The opening diameter of the iron wok is 30 cm, and the depth is 10 cm. As shown in Figure 6, the ultrasonic transducers are placed towards the center of the iron wok.

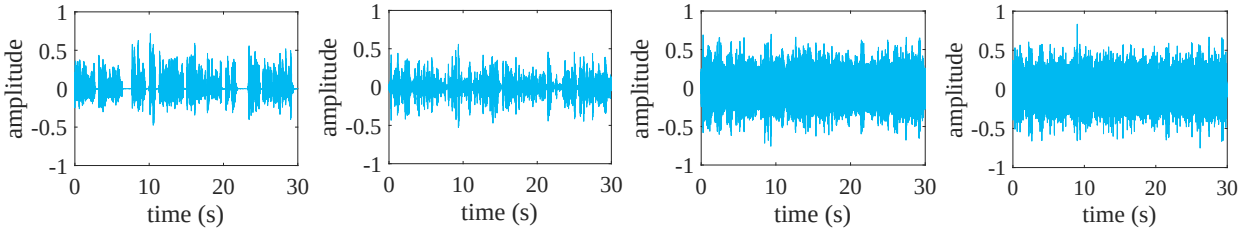
5.1.2 Format of Key. As we have mentioned in Section 4, Patronus uses the frequency sequence as the key. This key must include the duration of each frequency in addition to the frequency itself in order for the Descramble Receiver to generate the scramble waveform. Thus, our key file includes the frequency sequence plus the sample rate of the Scramble Transmitter and the number of samples of each frequency.

5.2 Descramble Receiver for Authorized Devices

We use an ordinary smartphone as an authorized device. The authorized device receives the key from the Scramble Transmitter. After the audio is recorded, the smartphone reconstructs the scramble waveform with the given key and leverages NLMS Adaptive filter to cancel the scramble. Formally, it takes the following steps:

5.2.1 Reconstruct Scramble Waveform. As we mentioned, in addition to the frequency sequence, the received key also contains the sampling rate of the Scramble Transmitter, which is denoted by f_{st}, as well as the number of samples of each frequency n_t. With the known sampling rate of the authorized device f_{sr}, the number of its recovered samples for each scramble frequency component can be calculated through the equation

$$n_r = \frac{f_{sr} n_t}{f_{st}}, \tag{5}$$



(a) Original waveform (b) Authorized waveform (c) Unauthorized waveform (d) Descrambled by STFT attack

Figure 7: Illustration of original waveform, authorized waveform, unauthorized waveform, and descrambled waveform by STFT attack.

After getting n_r, the authorized device uses the same process as the Scramble Transmitter to generate the scramble, *i.e.,* generating the discrete cosine signal with the frequency f_i and f_{i+1}, and connecting them by a chirp signal with start frequency f_i and end frequency f_{i+1}, where f_i and f_{i+1} are from the frequency sequence in the key.

5.2.2 Normalized Least-Mean-Square (NLMS) Adaptive Filter. After reconstructing the scramble waveform, we can use the Normalized Least-Mean-Square Adaptive Filter to cancel the scramble from the scrambled record. Specifically, we put the scrambled record rec_s and the scramble waveform s into the NLMS Adaptive Filter to get the descrambled waveform e by removing s from rec_s. According to the discussion in Section 3, the scramble wave is not only generated by frequencies in the given frequency sequence but also generated by high-order frequencies that are multiples of the target frequencies. Therefore, after getting e from the NLMS Adaptive filter, we still need to iteratively remove the multiples of the frequency sequence scramble by NLMS Adaptive filter. It means that we iteratively put e and the scramble waveform generated by k-times multiple of the frequency sequence into NLMS Adaptive Filter, where $k = 2, 3, 4, 5, 6$ in our prototype.

In summary, the procedure of authorized devices for removing the scramble from the record is shown in Algorithm 1.

Algorithm 1 Remove Scramble from the record

Input: rec_s, f_{sr}, f_{st}, n_t,
the frequency sequence $f[1..n]$
Output: Speech Record without Scramble e

$n_r \leftarrow f_{sr} n_t / f_{st}$
$e \leftarrow rec_s$
for $k = 1$ to 6 **do**
  $s \leftarrow$ ScrambleGenerator($k \times f[1..n]$, n_r).
  $e \leftarrow$ NLMS-Adaptive-Filter(e, s)
return e

The NLMS-Adaptive-Filter can be found in many open-source libraries, *e.g.,* MATLAB, Python, *etc.*. Due to the selective frequency response of different smart devices, each model has its own parameter setting. In the implementation, we choose 500 as the number of taps and 0.005 as the step size for an iPhone, 100 as the number of taps and 0.003 as the step size for a Pixel, and 300 as the number of taps and 0.005 as the step size for a Galaxy S9.

5.3 Simulated STFT Attacker

We also simulate an STFT attacker to verify whether or not Patronus can prevent such an attack. Specifically, as discussed in Section 4.2.1, we apply STFT to the scrambled recording using the MATLAB function *stft* to infer its frequency sequence. We then feed the frequency sequence to an NLMS adaptive filter to get the descrambled recording. Experiment results are shown in Section 6.8. Here, we illustrate an example, which contains the original waveform, authorized waveform, unauthorized waveform and the waveform descrambled by STFT, in Figure 7. As illustrated by the figure, we observe that the authorized waveform is similar to the original waveform, the unauthorized waveform is different from the original one, and the unauthorized waveform is similar to the waveform descrambled by STFT attack. Therefore, our prototype proves that Patronus can block the unauthorized recording while allowing authorized recording, and it can prevent STFT attacks.

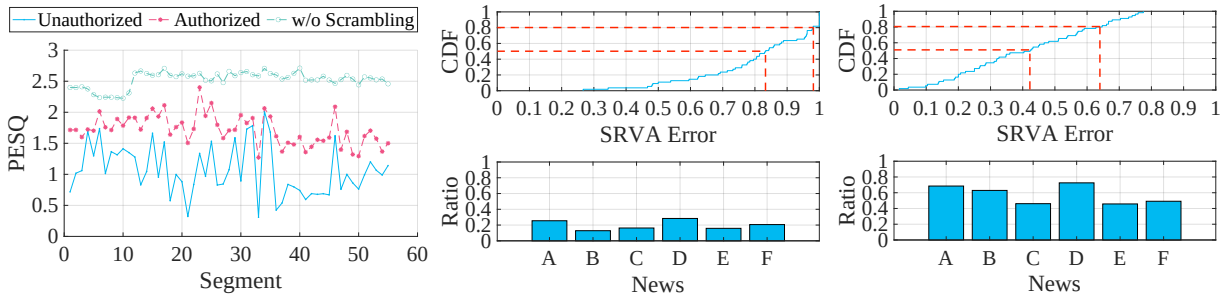
6 EVALUATION

6.1 Overview

To evaluate the performance of Patronus, we select six news speech waveforms from Voice of America (VOA) and note these waveforms as A - F. The news speeches are read by a male, a female, or both alternatively, sometimes with background music.

A normal speaker (shown in Figure 6) is set to play these news waveforms, and we also read the news ourselves. While the news waveforms are played under different conditions, we start Patronus to interfere with the unauthorized recording device. Meanwhile, an authorized device is recording too. Later we apply scramble cancellation to recordings from the authorized device. After getting the scrambled recordings and scramble-canceled recordings, the following metrics are adopted to measure the performance of Patronus.

6.1.1 Perceptual Evaluation of Speech Quality (PESQ). PESQ is a common-used metric of speech quality [15]. It is widely adopted by phone manufacturers, network equipment vendors, and telecom operators. Technically, the inputs include a clear speech signal as the reference and a signal that needs to be measured. The output is a Mean Opinion Score (MOS) [28] ranging from -0.5 to 4.5. A high PESQ score means that the corresponding speech has a high hearing quality and vice versa. Typically, PESQ values ranging from 1.00 to 1.99 means "No meaning understood with any feasible effort" while those ranging from 3.80 to 4.50 meaning "Complete



(a) PESQ (b) Scrambled (c) Descrambled

Figure 8: (a) PESQ of recordings captured by unauthorized and authorized devices, and PESQ of recordings without scrambling by turning off Patronus as the baseline. (b) Upper half: The CDF of SRVA Error of scrambled recordings from the unauthorized device. Lower half: The ratio of SRVA between scrambled recordings and original waveforms. (c) Upper half: The CDF of SRVA Error of descrambled recordings from the authorized device. Lower half: The ratio of SRVA between descrambled recordings and original waveforms.

relaxation possible; no effort required" [29]. However, we cannot regard the audio recording as strict as lossless communication. To fit PESQ to characterize the performance of Patronus, we measure the PESQ of recordings without scrambling by turning off Patronus, and use that result as the baseline. As shown in Figure 8(a), such recordings have PESQ between 2.2 and 2.7. We regard them as the upper bound of both unauthorized and authorized recordings. In the following experiments, we use the PESQ implementation written in MATLAB [30] to compute the PESQ score.

6.1.2 Speech Recognition Vocabulary Accuracy (SRVA). We also use a Speech Recognition service to measure the effectiveness of scrambling and descrambling. Specifically, we apply Google's Speech To Text (STT) service to transform the acoustic signals to text. We first use the STT service to recognize the original speech without interference and treat the recognized word sequence w_c as the ground truth. Then we use the STT service to recognize the scrambled speech and descrambled speech, and use w_s and w_d to denote their results, respectively. We name $\frac{\sum_{i \in w_s} isTrue(i \in w_c)}{|w_c|}$ (or $\frac{\sum_{i \in w_d} isTrue(i \in w_c)}{|w_c|}$) as the Speech Vocabulary Recognition Accuracy (SRVA) and use it to quantify the effectiveness of scrambling and descrambling. Note that $isTrue(i \in w_c)$ returns 1 when i is a word from w_c, and 0 when i is not a word from w_c. We define SRVA Error as 1 − SRVA which indicates the error rates of recognition with the STT service.

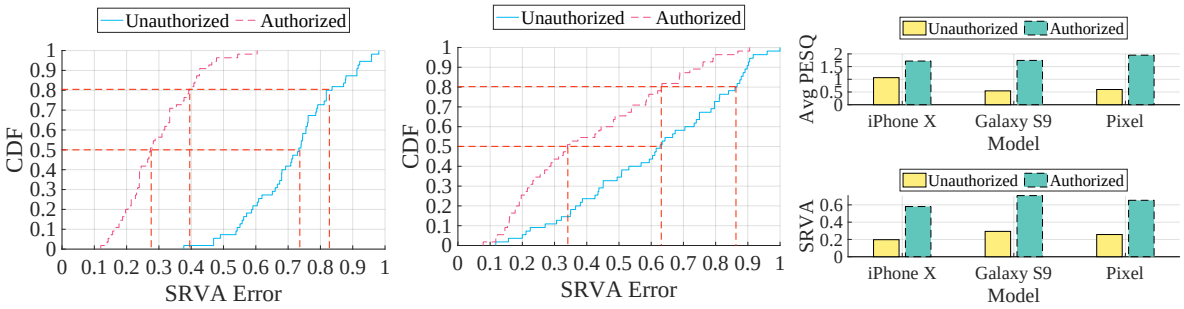
Using the above metrics, we try to answer the following questions:

- Can Patronus effectively scramble the unauthorized speech recordings?
- Can Patronus permit authorized devices to record the speech?
- Can Patronus work on different mobile devices?
- What is the impact of the distance between Patronus and a recorder?
- What is the impact of the reflection layer?
- What is the impact of the frequency switching time?
- Is it possible to perform real-time descrambling?

6.2 Effectiveness of Scrambling and Descrambling

We split the 6 news speech waveforms into 55 segments (1650 seconds in total), each 30 seconds long. Both the authorized and unauthorized device are Apple iPhone X in this experiment, so do the following experiments except that of Section 6.5. As shown in Figure 8(a), with Patronus's scrambling, the hearing qualities of most segments are extremely low. Specifically, 44 out of 55 (80.0%) segments have PESQ scores lower than 1.5. For SRVA, overall, only 551 out of 2796 (19.7%) words are recognized correctly. More detailed results are shown in Figure 8(b). The upper half shows the CDF of the SRVA Error. We can know that 50% of the recordings have SRVA Error lower than 0.84, and 80% of the recordings have SRVA Error lower than 0.98. The lower half shows the ratio of SRVA between scrambled recordings and original waveforms. The results show that all of the news waveforms having a recognition rate lower than 0.3. Here we want to mention that if a word appears multiple times in a speech, SRVA would result in a high value or a low value compared to the actual word recognition rate. However, duplicated words have little impact because the duplicate rates of every segment, *i.e.,* the ratio between the count of a specific word and the total count of words in the segment, are lower than 5%.

To evaluate the effectiveness of descrambling, an authorized device records the speech under the scrambling from Patronus. The authorized device then cancels the scramble using the received key. As shown in Figure 8(a), after descrambling, only 9 out of 55 (16.3%) segments having PESQ scores lower than 1.5. On average, descrambled recordings have 1.6x higher PESQ scores than their corresponding scrambled recordings. As for SRVA, we show the CDF of the SRVA Error in the upper half of Figure 8(c). These results show that 50% of the descrambled recordings have SRVA Error lower than 0.43, which is 49% lower than scrambled recordings. Moreover, 80% of the descrambled recordings have SRVA Error lower than 0.64, which is 35% lower than scrambled recordings. As shown in the lower half of Figure 8(c), ratios of SRVA between



(a) Human reading (b) Human hearing (c) Different models

Figure 9: (a) Compare SRVA between before and after descrambling for the human voice. (b) Compare SRVA between before and after descrambling for human recognition. (c) Compare average PESQ and SRVA among different models.

descrambled recordings and original waveforms are higher than 0.4 and lower than 0.8. They are at least 2x better than the scrambled recordings. The quality of the descrambled recordings is not as good as the original ones because there are residual components of the scramble after applying the NLMS adaptive filter. Moreover, background music and the volume of the original waveform also affects the quality of the descrambled recordings. For example, news C has a lower ratio after being descrambled by the authorized device compared to the other news clips because it has background music that could affect the performance of authorized devices. It also affects the SRVA of the record without scrambling, i.e., only 223 words are recognized from 295 in total. The reader of news E reads the news in a lower volume compared to others, so it has a lower ratio after being descrambled by the authorized device compared to the other news clips.

6.3 Effectiveness of Human Voice Scrambling and Descrambling

To verify whether Patronus works for real human speaking other than a sound player, we read the news and calculate SRVA[1]. As shown in Figure 9(a), Patronus can effectively scramble and descramble the human voice. Specifically, for the scrambled recordings, the median of SRVA Error is 0.74, and 80% of scrambled recordings have SRVA Error lower than 0.83. For the descrambled recordings, the median of SRVA Error is 0.27, and 80% of the descrambled recordings have SRVA Error lower than 0.4. The descrambling effectiveness of the human speaker is better than that of recorded sounds because recorded sounds from VOA sometimes play background music.

6.4 Effectiveness of Human Recognition to Scrambled Recordings and Descrambled Recordings

Because there might exist differences between machine learning-based speech recognition and human speech recognition, we invite 11 volunteers to write down words after listening to the 55 scrambled recordings and 55 descrambled ones[1]. The results are shown in Figure 9(b). People react differently to noise. Some people are very sensitive and the scrambled noise make them very uncomfortable. Note, the noise is generated by ultrasound speakers and only captured by the nonlinear effects of microphones, so it will not disturb the people in the original conversation. It will only be heard after getting recorded by unauthorized devices. Further, authorized devices will be able to filter out such noises eliminating the discomfort for those listeners. The recovered information from humans listening to descrambled recordings is still better than that of humans listening to scrambled ones. 50% of the scrambled recordings have SRVA Error lower than 0.63, and 80% of the scrambled recordings have SRVA Error lower than 0.86. As a comparison, 50% of the descrambled recordings have SRVA Error lower than 0.34, and 80% of the descrambled recordings have SRVA Error lower than 0.63.

6.5 Effectiveness on Different Mobile Models

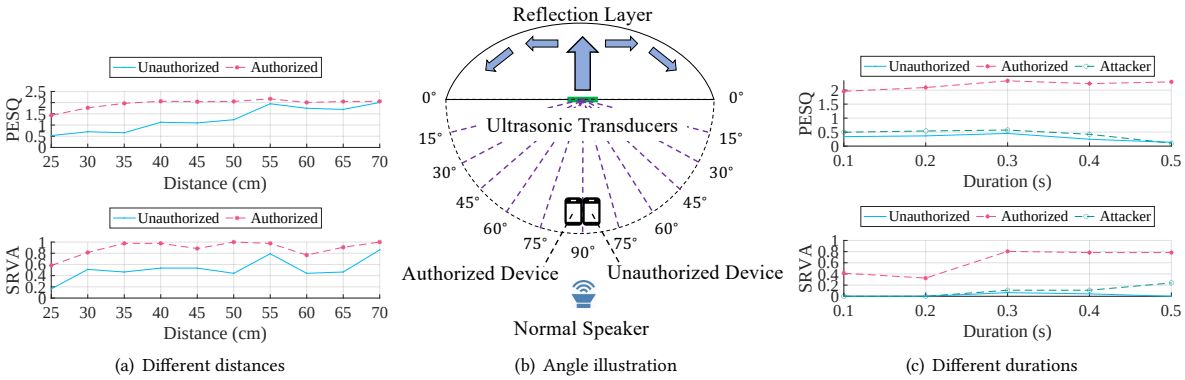
To verify whether Patronus works on different mobile models, we test it on three devices, an Apple iPhone X, a Samsung Galaxy S9, and a Google Pixel. We play all 55 segments using the normal speaker, and calculate average PESQs and SRVAs.

As shown in Figure 9(c), less than 30% of words can be recognized by the STT service for all the unauthorized devices, and around 65% of words can be recognized for all the authorized devices. When the mobile devices are unauthorized, the average PESQ of iPhone X is 1.06, and the average PESQ of the other two models are even lower, roughly 0.5. When the mobile devices are authorized, they all achieve an average PESQ around 1.85. This demonstrates that Patronus works well for all devices; namely, it prevents all models from making good unauthorized recordings and allows all models to make acceptable authorized recordings.

6.6 Impact of the Distance

We also characterize the impact of the distance between Patronus and the recording devices (both authorized and unauthorized). We put the Scramble Transmitter at the origin. A randomly-picked speech segment (which has 43 words) is played by a normal speaker, which simulates the talker. The authorized device and an unauthorized device are recording at the same time. Their distance to the Scramble Transmitter varies from 25 cm to 70 cm. Results of SRVA

[1]This experiment is approved by IRB committee.



(a) Different distances (b) Angle illustration (c) Different durations

Figure 10: (a) Compare PESQ and SRVA at different distances. (b) Illustration of the reflection layer experiment. (c) Compare PESQ and SRVA with different frequency switching times.

DT (ms) \ MSO / RT (s)	1	2	3	4	5	6
1	51	96	159	209	265	328
2	73	145	218	291	373	454
5	161	322	487	634	798	954
10	290	582	851	1108	1389	1653
20	548	1094	1653	2165	2695	3298
30	822	1617	2348	3088	3830	4563

Table 1: Descramble time (DT) of different record times (RT) with different max scramble orders (MSO, the upper bound of k in Algorithm 1).

and PESQ between two devices are shown in Figure 10(a). Overall, as the distance increases, the ultrasound would attenuate more. Therefore, the strength of the scramble decreases as the distance from the scramble transmitter increases. As a result, when the device is far enough away, both the authorized and unauthorized device can both record a clear speech. On the other hand, when devices are close enough, unauthorized devices produce recordings that are severely scrambled whereas authorized devices can recover much clearer speech using the secret key. The working area can be extended by using high power ultrasonic speakers, which we will discuss later. Here we want to mention that although there is a bump in Figure 10(a) at 55 cm with the SRVA, PESQs of 55cm and 60cm are close. This means that humans cannot see much difference between these two recordings, something we confirmed in person by listening to these recordings with this objective in mind. Thus, the SRVA bump at 55cm might be due to an error-correction mechanism of the Google STT engine; of course, since this is proprietary technology, we do not know how or why this error-correction would produce such a performance bump for this recording.

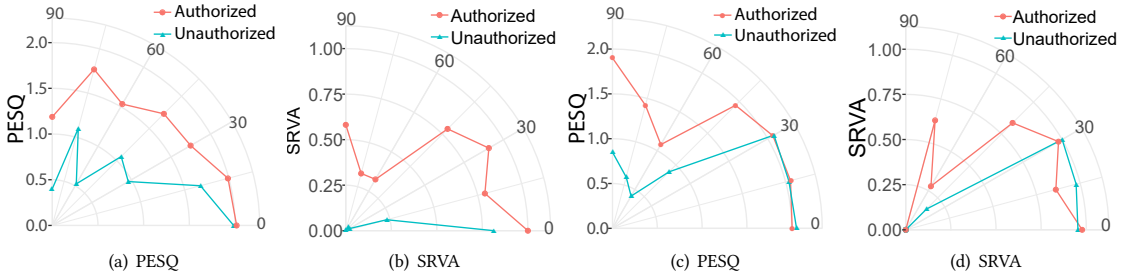
6.7 Impact of the Reflection Layer

As we mentioned before, the ultrasound wave often propagates along a straight line. To enlarge the range of Patronus scrambling, we design a reflection layer. In this experiment, we apply the common speaker to play the chosen speech segment (43 words). As shown in Figure 10(b), we point the ultrasonic transducers towards the reflection layer and change angles of both authorized and unauthorized devices to the ultrasonic transducers and measure Patronus' performance; in other experiments, the devices are always put at the $90°$ angle. We also measure the performance without using the reflection layer. We turn the ultrasonic transducers around so they face in the same direction as the normal speaker when we remove the reflection layer. The results when using the reflection layer are shown in Figure 11(a) and 11(b), and the results without using the reflection layer are shown in Figure 11(c) and 11(d). From the results, we see that with the reflection layer, Patronus can successfully scramble the unauthorized device when the angle is more than $15°$, which is signficantly larger than the angle of more than $45°$ needed by Patronus without the reflection layer. Therefore, the reflection layer does significantly enlarge the scramble range of Patronus.

6.8 Impact of the Frequency Duration

We also measure the impact of the frequency duration. As we discussed in Section 4, we would like to make the duration of each frequency as short as possible. However, the shorter the frequency duration is, the harder it is for authorized devices to descramble. To verify this feature, we put an authorized and an unauthorized device at 40 cm to Patronus and play the chosen segment (43 words) using the normal speaker. Both devices record the speech under Patronus using 5 different frequency durations: 0.1 s, 0.2 s, 0.3 s, 0.4 s and 0.5 s. We calculate PESQs and SRVAs for each duration. Moreover, we implement the attack model from Section 4.2, which first calculates approximate scramble frequencies using STFT and then attempts to cancel the scramble using an NLMS adaptive filter. We calculate PESQs and SRVAs for each duration and all devices including the attack model.

As shown in Figure 10(c), for all durations, SRVAs of the unauthorized device are lower than 0.1, and PESQs are lower than 0.5. The authorized device has higher SRVAs and PESQs than the unauthorized device. Specifically, when the duration comes to 0.3 s, the SRVA reaches roughly 0.8 and PESQ exceeds 2.0. This verifies our claim that authorized devices can successfully descramble when the frequency duration is long enough.



(a) PESQ (b) SRVA (c) PESQ (d) SRVA

Figure 11: (a) and (b): Compare PESQ and SRVA with the using of the reflection layer. (c) and (d): Compare PESQ and SRVA without the using of the reflection layer.

A shorter duration also makes it harder for attackers to crack the scrambled record, *e.g.,* SRVAs for the attacker also increase as the duration increases. Although both SRVAs and PESQs are higher than those of the unauthorized device, they are still too low to extract useful information. The reason why the NLMS adaptive filter fails is that the attacker cannot identify the scramble frequencies with enough accuracy. NLMS adaptive filter solves the optimization problem defined by Equation (4), which estimates the weight vector h_2'. Since convolution does not change the frequency of the signal, the attacker cannot make up for any offset existing between the correct frequency and the result from STFT. According to the frequency resolution problem of STFT as discussed in Section 4.3.4, the simulated attacker in our experiment gets an average frequency offset around 3 Hz, which makes it hard to descramble the recording.

6.9 Descramble Time

Sometimes when we grant recording permission to a specific speaker, the speaker would like to perform real-time descrambling. Patronus can achieve this working with real-time smart devices such as Amazon Alexa. To prove this, we measure the descramble time for records with different durations on a laptop with an Intel Core i7-4870HQ 2.5 GHz CPU. Since different high-order scramble waves (second-order component, third-order component, ...) may exist in a record simultaneously, we measure descramble time as a function of different max scramble orders, *i.e.,* the upper bound of k in Algorithm 1. As shown in Table 1, Patronus can descramble the record quickly. Specifically, when the record time is 1 s, Patronus can finish descrambling in 328 ms, even when the max scramble order is 6. This means that Patronus supports real-time descrambling.

7 LIMITATIONS AND FUTURE WORKS

Range: In our implementation, we use cheap and low power ultrasonic transducers to build the Scramble Transmitter. The result is a short working distance, *i.e.,* less than 70 cm. To enlarge the working area to a wider range of angles, we designed a reflection layer and verified that it could enlarge the working area by using an iron wok in our prototype. We can also use a high power ultrasonic speaker to protect a larger area. Some commercial off-the-shelf devices can emit ultrasound which could be sensed in a larger area. For example, UPS+ [5] uses an ultrasonic speaker with a working area of 50m × 50m. However, it is expensive. We can reduce the cost by deploying one expensive speaker and multiple transducers like UPS+[5]. Here we provide users with three options to deploy Patronus according to their requirements such as working area and budget. The first option is to use cheap transducers and a reflection layer to protect a small area. The second is combining an expensive speaker and multiple transducers to protect a larger area. The third is using multiple expensive speakers to protect the largest area.
Volume: In our implementation, we assume the talker uses a normal volume, *i.e.,* not too loud or too quiet. However, the performance of Patronus does vary as a function of the speaker volume. For example, if the talker speaks too loudly, the scramble cannot mess up the recording; in the opposite extreme, a quiet talker cannot be recovered using descrambling. To adapt to different volumes, we can add a microphone to measure the talker's volume. With multiple deployed ultrasonic speakers or transducers, we can first detect the position of recording devices and then adjust the power of ultrasound emitted from the nearest speakers according to the talker's volume. There are two challenges that need to be solved. First, the microphone we use to measure the talker's volume can also be scrambled. Second, we need to localize recording devices before emitting scrambles. We leave these challenges as future work.

8 CONCLUSION

Acoustic privacy protection has always been an important topic. In this paper, we study the nonlinear effects on commercial off-the-shelf microphones. Based on our study, we propose Patronus, which leverages the nonlinear effects to disrupt unauthorized devices from recording the speech while simultaneously allowing authorized devices to record clear speech audio. We implement and evaluate Patronus in a wide variety of representative scenarios. Results show that Patronus effectively blocks unauthorized devices from making secret recordings while allowing authorized devices to successfully make clear recordings.

ACKNOWLEDGEMENT

We sincerely thank anonymous reviewers and shepherd for insightful comments to improve our work, and thanks to Dr. Eric Torng for the help of careful proofreading. This work is supported by the National Science Foundation under grant NSF 1919154.

REFERENCES

[1] The Guardian. Apple apologises for allowing workers to listen to siri recordings. https://www.theguardian.com/technology/2019/aug/29/apple-apologises-listen-siri-recordings. (Accessed on Feb. 28, 2020).

[2] CNBC. Amazon echo recorded conversation, sent to random person: report. https://www.cnbc.com/2018/05/24/amazon-echo-recorded-conversation-sent-to-random-person-report.html. (Accessed on Feb. 28, 2020).

[3] The Guardian. Ukraine prime minister offers resignation after leaked recording. https://www.theguardian.com/world/2020/jan/17/ukraine-prime-minister-oleksiy-goncharuk-offers-resignation-after-leaked-recording. (Accessed on Feb. 28, 2020).

[4] Yu-Chih Tung and Kang G. Shin. Exploiting sound masking for audio privacy in smartphones. In *Proceedings of ACM ASIACCS, July 7–12, 2019, Auckland, New Zealand.*

[5] Qiongzheng Lin, Zhenlin An, and Lei Yang. Rebooting ultrasonic positioning systems for ultrasound-incapable smart devices. In *Proceedings of ACM MobiCom, October 21–25, 2019, Los Cabos, Mexico.*

[6] Anti-eavesdropping and recording blocker device, China Patent 201320228440, Oct. 2013.

[7] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. Backdoor: Making microphones hear inaudible sounds. In *Proceedings of ACM MobiSys, June 19–23, 2017, Niagara Falls, NY, USA.*

[8] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. Inaudible voice commands: The long-range attack and defense. In *Proceedings of USENIX NSDI, April 9–11, 2018, Renton, WA, USA.*

[9] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. In *Proceedings of ACM CCS, October 30–November 3, 2017, Dallas, TX, USA.*

[10] Tao Chen, Longfei Shangguan, Zhenjiang Li, and Kyle Jamieson. Metamorph: Injecting inaudible commands into over-the-air voice controlled systems. In *Proceedings of NDSS, February 23–26, 2020, San Diego, CA, USA.*

[11] Xinyan Zhou, Xiaoyu Ji, Chen Yan, Jiangyi Deng, and Wenyuan Xu. Nauth: Secure face-to-face device authentication via nonlinearity. In *Proceedings of IEEE INFOCOM, April 29–May 2, 2019, Paris, France.*

[12] Qiben Yan, Kehai Liu, Qin Zhou, Hanqing Guo, and Ning Zhang. Surfingattack: Interactive hidden attack on voice assistants using ultrasonic guided wave. In *Proceedings of NDSS, February 23–26, 2020, San Diego, CA, USA.*

[13] Aleksandr Rovner. The principle of ultrasound. https://www.echopedia.org/wiki/The_principle_of_ultrasound, 2015. (Accessed on Oct. 19, 2020).

[14] Ali H Sayed. *Fundamentals of adaptive filtering*. John Wiley & Sons, 2003.

[15] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *Proceedings of IEEE ICASSP, May 7–11, 2001, Salt Lake City, UT, USA.*

[16] Yitao He, Junyu Bian, Xinyu Tong, Zihui Qian, Wei Zhu, Xiaohua Tian, and Xinbing Wang. Canceling inaudible voice commands against voice control systems. In *Proceedings of ACM MobiCom, October 21–25, 2019, Los Cabos, Mexico.*

[17] Anran Wang, Chunyi Peng, Ouyang Zhang, Guobin Shen, and Bing Zeng. Inframe: Multiflexing full-frame visible communication channel for humans and devices. In *Proceedings of ACM HotNets, October 27–28, 2014, Los Angeles, CA, USA.*

[18] Anran Wang, Zhuoran Li, Chunyi Peng, Guobin Shen, Gan Fang, and Bing Zeng. Inframe++ achieve simultaneous screen-human viewing and hidden screen-camera communication. In *Proceedings of ACM MobiSys, May 18–22, 2015, Florence, Italy.*

[19] Viet Nguyen, Yaqin Tang, Ashwin Ashok, Marco Gruteser, Kristin Dana, Wenjun Hu, Eric Wengrowski, and Narayan Mandayam. High-rate flicker-free screen-camera communication with spatially adaptive embedding. In *Proceedings of IEEE INFOCOM, April 10–15, 2016, San Francisco, CA, USA.*

[20] Kai Zhang, Chenshu Wu, Chaofan Yang, Yi Zhao, Kehong Huang, Chunyi Peng, Yunhao Liu, and Zheng Yang. Chromacode: A fully imperceptible screen-camera communication system. In *Proceedings of ACM MobiCom, October 29–November 2, 2018, New Delhi, India.*

[21] Qian Wang, Kui Ren, Man Zhou, Tao Lei, Dimitrios Koutsonikolas, and Lu Su. Messages behind the sound: real-time hidden acoustic signal capture with smartphones. In *Proceedings of ACM MobiCom, October 3–7, 2016, New York, NY, USA.*

[22] Man Zhou, Qian Wang, Kui Ren, Dimitrios Koutsonikolas, Lu Su, and Yanjiao Chen. Dolphin: Real-time hidden acoustic signal capture with smartphones. *IEEE Transactions on Mobile Computing*, 18(3):560–573, 2018.

[23] Lan Zhang, Cheng Bo, Jiahui Hou, Xiang-Yang Li, Yu Wang, Kebin Liu, and Yunhao Liu. Kaleido: You can watch it but cannot record it. In *Proceedings of ACM MobiCom, September 7–11, 2015, Paris, France.*

[24] Shilin Zhu, Chi Zhang, and Xinyu Zhang. Automating visual privacy protection using a smart led. In *Proceedings of ACM MobiCom, October 16–20, Snowbird, Utah, USA.*

[25] Ingo R Titze and Daniel W Martin. Principles of voice production, 1998.

[26] Ronald J Baken and Robert F Orlikoff. *Clinical measurement of speech and voice*. Cengage Learning, 2000.

[27] Sheng Shen, Nirupam Roy, Junfeng Guan, Haitham Hassanieh, and Romit Roy Choudhury. Mute: bringing iot to noise cancellation. In *Proceedings of ACM SIGCOMM, August 20–25, 2018, Budapest, Hungary.*

[28] ITUT Rec. P. 800.1, mean opinion score (mos) terminology. *International Telecommunication Union, Geneva*, 2006.

[29] Mika Wilson. Pesq - what is it and how could it transform your customer experience? https://www.spearline.com/blog/post/pesq---what-is-it-and-how-could-it-transform-your-customer-experience-/, 2018. (Accessed on Oct. 2, 2020).

[30] Kamil Wojcicki. Pesq matlab wrapper. https://www.mathworks.com/matlabcentral/fileexchange/33820-pesq-matlab-wrapper. (Accessed on Mar. 6, 2020).