Spying with Your Robot Vacuum Cleaner: Eavesdropping via Lidar Sensors

Sriram Sami
National University of Singapore
srirams@comp.nus.edu.sg

Yimin Dai
National University of Singapore
e0505408@u.nus.edu

Sean Rui Xiang Tan
National University of Singapore
seantanr@comp.nus.edu.sg

Nirupam Roy
University of Maryland, College Park
nirupam@cs.umd.edu

Jun Han
National University of Singapore
junhan@comp.nus.edu.sg

ABSTRACT

Eavesdropping on private conversations is one of the most common yet detrimental threats to privacy. A number of recent works have explored side-channels on smart devices for recording sounds without permission. This paper presents *LidarPhone*, a novel acoustic side-channel attack through the lidar sensors equipped in popular commodity robot vacuum cleaners. The core idea is to repurpose the lidar to a laser-based microphone that can sense sounds from subtle vibrations induced on nearby objects. *LidarPhone* carefully processes and extracts traces of sound signals from inherently noisy laser reflections to capture privacy sensitive information (such as *speech* emitted by a victim's computer speaker as the victim is engaged in a teleconferencing meeting; or known music clips from television shows emitted by a victim's TV set, potentially leaking the victim's political orientation or viewing preferences). We implement *LidarPhone* on a Xiaomi Roborock vacuum cleaning robot and evaluate the feasibility of the attack through comprehensive real-world experiments. We use the prototype to collect both spoken digits and music played by a computer speaker and a TV soundbar, of more than 30k utterances totaling over 19 hours of recorded audio. *LidarPhone* achieves approximately 91% and 90% average accuracies of digit and music classifications, respectively.

CCS CONCEPTS

• **Computer systems organization** → **Sensors and actuators**;
• **Security and privacy** → **Embedded systems security**.

KEYWORDS

Lidar, Acoustic Side-Channel, Eavesdropping

ACM Reference Format:
Sriram Sami, Yimin Dai, Sean Rui Xiang Tan, Nirupam Roy, and Jun Han. 2020. Spying with Your Robot Vacuum Cleaner: Eavesdropping via Lidar Sensors. In *The 18th ACM Conference on Embedded Networked Sensor Systems (SenSys '20), November 16–19, 2020, Virtual Event, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3384419.3430781

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

ACM ISBN 978-1-4503-7590-0/20/11…$15.00
https://doi.org/10.1145/3384419.3430781

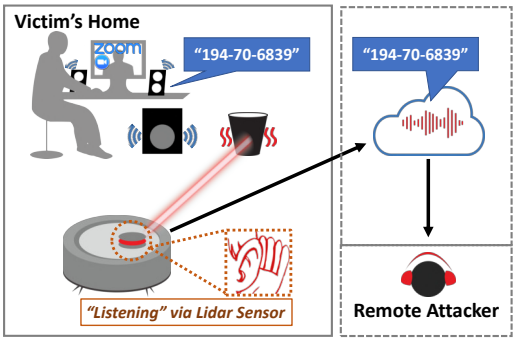

Figure 1: Figure depicts the *LidarPhone* attack, where the adversary remotely exploits the lidar sensor equipped on a victim's robot vacuum cleaner to capture parts of privacy sensitive conversation (e.g., credit card, bank account, and/or social security numbers) emitted through a computer speaker as the victim engages in a teleconference meeting.

1 INTRODUCTION

The proliferation of smart sensing devices in our homes opens up many opportunities for acoustic side-channel attacks on private conversations. Recently a number of academic papers reveal loopholes in smartphone sensors that potentially allow conversations to be recorded without permission [2, 21, 40, 54, 80]. Voice activated devices, smart speakers, and smart security cameras are often considered as sources of potential privacy threats. While devices with sound recording interfaces are the usual suspects, this paper launches a novel acoustic side-channel attack from a seemingly innocuous household device – a vacuum cleaning robot. Many of these indoor robots are equipped with lidars – a laser-based sensor for navigation [55, 77]. We develop a system to repurpose the lidar sensor to sense acoustic signals in the environment, remotely harvest the data from cloud, and process the raw signal to extract information. We call this eavesdropping system *LidarPhone*.

Sounds are essentially pressure waves that propagate through the vibrations of the medium. Hence, sound energy in the environment is partially induced on nearby objects creating subtle physical vibrations within those solid media. The fundamental concept of *LidarPhone* lies in sensing such induced vibrations in household objects using the vacuum robot's lidar sensor and then processing the recorded vibration signal to recover traces of sounds. This sensing method is inspired by the principles of *laser microphones* that

use reflected laser beams to sense sounds from vibrating objects. Although *laser mics* require sophisticated setups, the rotating lidar sensors are equipped with at least a laser transmitter and reflection sensor. This enables the key possibility to transform a lidar to a microphone. Figure 1 illustrates a potential attack scenario, where the adversary launches a remote software attack on the vacuum cleaner (as often witnessed in recent security breaches [19, 22, 47, 49]) to obtain lidar sensor readings. Of course, a practical implementation of this idea requires overcoming several challenges.

One of the main challenges in using a lidar as a *laser mic* is the extremely *low signal-to-noise ratio (SNR)* of the reflected signals. This is in part due to *LidarPhone*'s reliance on different physical principles than *laser mics*, despite their apparent high-level similarities. *Laser mics* target highly reflective materials (i.e., producing *specular reflections*) such as glass windows, which when vibrating cause significant changes to the return path and/or focus of the reflected laser, leading to high SNR. By contrast, a lidar's hardware amplifiers and analog-to-digital converter (ADC) are tuned to be sensitive *only to low intensity signals* as they are mostly reflecting off of rough surfaces such as trashcans thereby producing *diffuse reflections*. Hence, even if the lidar receives high intensity signals from a glass window, it would saturate its receiver and provide no useful information. Furthermore, the lidar has a fixed receiver at an adjacent position to its transmitter, making it difficult to receive specular reflections, as they are only reflected off of the glass at one particular angle. This is why a *laser mic* requires the adversary to manually align its receiver's position accordingly.

Furthermore, the sound signals are attenuated as the objects are not in contact with the speaker (i.e., *mechanically decoupled*). Also, the minutely vibrating objects attenuate some frequency components of the signal while adding additional noise. To overcome this challenge of low SNR, we utilize different signal processing techniques including *filtering* of the frequency components that contain noise. To further reduce the effect of noise, we perform *noise reduction* by subtracting the dynamically profiled noise using spectral subtraction [5]. Moreover, we *equalize* the signal by increasing the gain of (i.e., "boosting") the lower frequency components, as high frequency components of the signals are attenuated by the objects.

The other major challenge in designing *LidarPhone* attack is due to the lidar's low sampling rate. Given its rotating motion, the sampling rate for a single point on a target object is equivalent to the lidar's *rotation frequency*. We further increase the sampling rate by considering an attack when the lidar is put to halt to *rotation frequency* (often 5 Hz) × *samples per rotation* (typically 360), which increases the sampling rate from 5 Hz to 1.8 kHz for the case of a Xiaomi Roborock vacuum cleaner lidar. Despite the large improvement, 1.8 kHz is still significantly below the minimum frequency of 5 kHz for obtaining an intelligible speech signal [51]. Hence, we utilize supervised learning techniques by extracting relevant features to classify a list of digits, perform speaker and gender identification, and infer known sound clips played during TV shows. We leverage deep learning techniques through our use of convolutional neural networks. The seemingly innocuous information extracted by our model may leak privacy sensitive information including credit card, bank account, and/or social security numbers, as well as the victim's political orientation from news introduction music.

The vibration sensing mechanism and sound inference techniques are core to *LidarPhone*. Additionally, we build on existing reverse engineering building blocks to demonstrate a proof-of-concept remote system hijack that allows attackers to control the robot and capture sensor data. We implement a prototype of *LidarPhone* on a Xiaomi Roborock vacuum cleaning robot and evaluate the feasibility of the attack through comprehensive real-world experiments. We collect digit utterances and music excerpts played with a computer speaker and a TV soundbar by pointing the lidar at several common household objects (including trash cans or paper bags), collecting 30k utterances totaling over 19 hours of recorded audio. From our empirical analysis, *LidarPhone* achieves digit and music classification accuracies of 91% and 90%, respectively. Overall, we make the following contributions:

- We introduce a novel *remote eavesdropping attack* that utilizes lidar sensors on commodity robot vacuum cleaners.
- We present the design and implementation of *LidarPhone* by introducing and solving corresponding challenges inherently surpassing those of existing *laser mics*.
- We evaluate *LidarPhone* with real-world experiments using a commodity robot vacuum cleaner, a computer speaker, and a TV soundbar to demonstrate its feasibility.

Through this work, we reveal a new direction of side-channel attacks that exploits active light sensors. While we investigate lidars on a vacuum cleaner as an example, our findings may easily be extended to many other active sensors including smartphone infrared sensors (for face recognition [75]) and motion detector PIR sensors [6, 32]. We hope that our findings will spur research on detecting forthcoming attacks and new defense methods.

2 A PRIMER ON LASER-BASED SENSING

We present the relevant background information on lidar sensors and laser-based microphones.

2.1 Lidar Sensor

A Light Detection and Ranging (lidar) sensor is designed to scan the surrounding scene by utilizing laser-based ranging techniques to create a distance map. Specifically, the lidar steers an infrared laser beam toward a target and measures the time-delay of the reflected beam to estimate the time-of-flight (Δt) of the signal. With the known speed of the laser signal (C), the distance to the target object (d) is measured through a simple calculation ($d = \frac{C \times \Delta t}{2}$) [78]. In practice, however, the physical method to measure distance varies depending on the accuracy, resolution, range of operation, and complexity of the electro-mechanical lidar sensor. Low-cost sensors (e.g., in robot vacuum cleaners) adopt a geometric approach that estimates distances from the angles of the transmitted and reflected beams [29] as illustrated in Figure 2.

The recent popularity of lidar sensors is due to their iconic presence in autonomous vehicles [14]. More recently, lidars are also frequently used in many commodity vacuum cleaning robots [11–13, 34, 44–46, 53]. They utilize lidar sensors for navigation and mapping purposes as they clean houses. The pervasiveness of such lidar sensors opens up avenues for scalable attack opportunities for adversaries, that we demonstrate through *LidarPhone*.

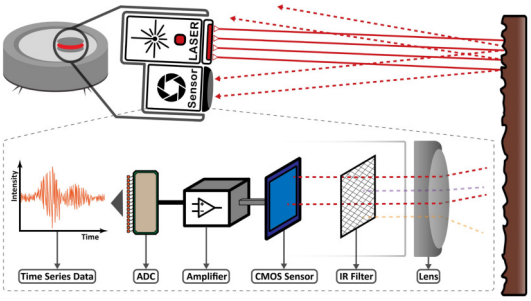

Figure 2: Operating principles of an inexpensive lidar: light reflected from a surface is focused through a lens, non-infrared frequencies are removed, the laser signal is captured by an imaging sensor, then amplified and quantized to create the final signal.

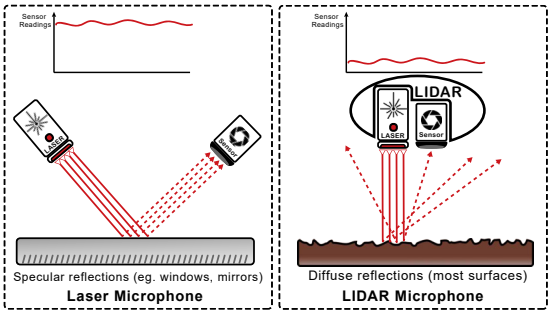

Figure 3: Differences between focused specular reflections relied upon by a *laser mic* (left) versus lower SNR diffuse reflections received by *LidarPhone* (right).

2.2 Laser Microphones and Their Limitations

The coherent signal source of lasers and their small wavelength (a few hundred nanometers) enable fine-grained distance measurement, which can be utilized to measure subtle motions or vibrations. This property of lasers led to a technique for long-range audio eavesdropping, namely the *laser microphone* [41]. Sound travels through a medium as a mechanical wave and induces minute physical vibrations in nearby objects. The key function of a *laser mic*, often used a spying tool, is to shine a laser beam on an object placed close to the sound source and measure this induced vibration to recover the source audio. A *laser mic* pointed at a glass window of a closed room can reveal conversations from inside the room from over 500 meters away [57]. To achieve high intensity reflections, *laser mics* require the adversary to manually align the transmitter and the receiver to obtain **specular reflections** of light from highly-reflective surfaces (e.g., glass). Unlike in **diffuse reflections**, almost all incoming light energy is returned at a specific angle (i.e., angle of incidence) as depicted in Figure 3. Hence, the intensity amplitude of specular reflections is significantly higher than that of diffuse reflections.

However, the inexpensive lidar sensors equipped by the robot vacuum cleaners are manufactured to operate **only on diffuse reflections** [30, 31]. This is appropriate for navigation in an environment where most reflecting surfaces are not smooth, and the sensor must function with a fraction of the returned light intensity of a specular reflection. This phenomenon poses a significant challenge to *LidarPhone*'s design as the diffuse reflections significantly limit the available signal-to-noise-ratio (SNR) for high resolution vibration detection as depicted in Figure 3. Since the hardware amplifiers and ADC are optimized only for low amplitude diffuse reflections, our preliminary experiments show that even if the lidar sensor opportunistically finds a position (near a glass window or a mirror) to receive a high intensity specular reflection, the received signals are saturated and clipped. To successfully sense sounds using lidar sensors, *LidarPhone* must operate with low-intensity signals, and recover signals that are close to the noise floor.

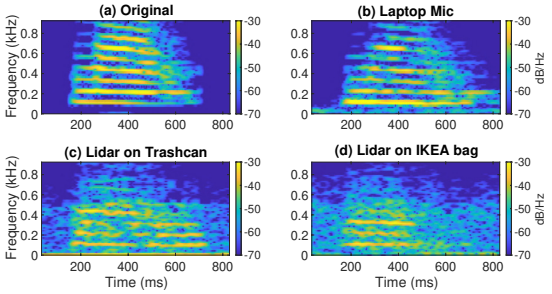

Figure 4: Figure depicts the utterance ("nine") captured by: (a) original audio (b) microphone recording at 3 m (c) processed lidar recordings from reflections off of a trashcan and (d) an IKEA bag. The figures illustrate the feasibility of capturing speech signals with a lidar. The corresponding audio files are available at bit.ly/lidarphone-sensys.

3 FEASIBILITY STUDY

We present a preliminary study demonstrating *LidarPhone*'s feasibility by playing an utterance of the digit *"nine"* through a computer speaker and recording it with a laptop microphone and a lidar. The lidar records the sound by capturing the laser reflections off of two objects positioned near the speaker – a trashcan covered with a translucent plastic trash bag, and a polypropylene IKEA plastic bag.

Figure 4 illustrates the spectrogram of the four signals, namely (a) original, (b) laptop microphone, (c) lidar recordings reflected off of the trashcan, and (d) the IKEA bag. The spectrograms depict the corresponding frequency (*kHz*) with varying time (*ms*). The sampling rates of both the (a) original and (b) laptop microphone are 8kHz, while the two processed lidar recordings from both the (c) trashcan and (d) IKEA bag are 1.8kHz. We plot them until 0.9kHz for all four plots for consistency. From this study, we observe that the lidar is able to capture the speech signal, but with significantly reduced SNR. We also observe additional challenges of *LidarPhone*– that the lidar primarily captures only lower frequency components up to around 0.6kHz. This may be because the vibrating objects attenuate high frequency components as well as adding additional noise. Furthermore, we also observe that the SNR depends on the object's material. The IKEA bag is sturdier than the trashcan's plastic covering, and is therefore more difficult to deform due to incident acoustic energy; the spectrogram depicts the significant attenuation of the signal.

4 THREAT MODEL

We present *LidarPhone*'s threat model, namely the attacker's goal and capabilities, and further outline its assumptions.

Goal and Capabilities. The *goal* of the attacker is to launch a stealthy *remote eavesdropping attack* by utilizing the lidar readings from the victim's robot vacuum cleaner. The lidar captures sound signals by obtaining its reflections off of objects that minutely vibrate due to nearby sound sources (e.g., victim's computer speaker or TV soundbar). To achieve this goal, the attacker may launch two types of attacks, namely *Speech-based* and *Sound-based Attacks*. The attacker has *capabilities* to remotely exploit vulnerabilities in robot cleaners, often witnessed in recent real-world attacks [10, 26, 38, 70], to (1) stop the lidar spinning to capture reflections off of a single point of the object; and (2) obtain the corresponding raw lidar intensity values. Furthermore, the attacker has additional capabilities for each attack.

When launching a *Speech-based Attack*, the attacker targets potentially privacy-sensitive information from speech emitted by the computer speakers as the victim engages in a teleconferencing. There are three types of *Speech-based Attacks*: (1) *Digit Inference*, that predicts the spoken digit (out of a list of potential digits) to leak sensitive information including credit card, social security, and bank account numbers; (2) *Gender Inference*, that determines whether the speaker is a male or a female; and (3) *Speaker Inference*, that determines the identity of the speakers (out of a list of potential speakers). *Digit Inference* is a *targeted attack*, where the attacker targets a specific victim. Hence, the attacker has the capabilities to train on the targeted victim's recordings before launching this attack. For example, the victims may be high value targets such as political figures or celebrities, enabling the attacker to easily obtain labeled training data from publicly available recordings.

Second, when launching a *Sound-based Attack*, the attacker targets the introductory music of news programs emitted by the victim's TV soundbar, as different news channels exhibit certain political biases [3]. Hence, inferring the news channel that the victim watches on a regular basis may likely reveal his/her political orientation [73], often valued as privacy sensitive information [23].

Assumptions. We make the following assumptions when designing *LidarPhone*. First, we assume that the victim has a robot vacuum cleaner that is equipped with a lidar in his/her home or office. Second, we also assume that the victim has a commodity computer speaker and/or a TV soundbar along with everyday objects (e.g., trashcan or takeaway bag) positioned relatively near (i.e., within a few meters of) the victim's desk or TV stand. We discuss how the attacker identifies and targets these objects in Section 7.2.

5 ATTACK DESIGN AND IMPLEMENTATION

We present the details of *LidarPhone*'s design and implementation.

5.1 Design Overview

We present the modules that constitute the design of the two types of *LidarPhone* attack in Figure 5, namely the *Speech-* and *Sound-based Attacks*. The *Speech-based Attack* is divided into two phases, namely the *Bootstrapping* and *Prediction Phases*. In the *Bootstrapping Phase*, the attacker collects multiple acoustic signals captured by the lidar as training data. All of these data are first input to its

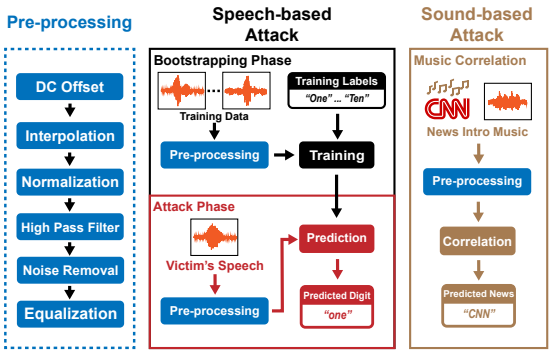

Figure 5: Figure depicts the design overview of *LidarPhone*. The captured audio signal is pre-processed in stages, and used to train models for digit, speaker or gender inference. In the attack stage, the recovered signal is tested against the appropriate model, or if performing a news music inference attack, used in a correlator module.

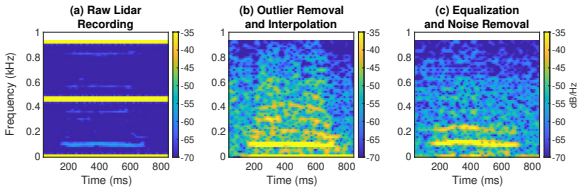

Figure 6: The effect of pre-processing stages on a *LidarPhone*-recorded utterance ("2"). The low SNR raw signal (a) has some information restored through outlier removal and interpolation to produce (b), which has equalization and noise removal applied to produce (c).

Pre-processing module to increase the signal-to-noise ratio (SNR). Subsequently, the pre-processed signals are input to the *Training* module along with their corresponding ground truth labels. In the *Attack Phase*, the attacker takes as input the lidar signal captured from the victim's home and applies the same pre-processing techniques before it is input to its *Prediction* module, along with the trained model from the *Bootstrapping Phase*. Ultimately, the *Prediction* module outputs the predicted speech such as spoken digits of credit card, bank account, and/or social security numbers, which was originally emitted from the victim's speakers.

Furthermore, the *Sound-based Attack* enables the attacker to infer the TV news channel that the victim watches by pattern matching the introduction music recorded by the lidar across a list of introduction music clips from popular news channels. The attacker initiates this attack by also applying the aforementioned pre-processing pipeline, which subsequently gets input to the *Correlation* module. In this module, the attacker correlates the input signal across the potential list of music clips and outputs the most likely one.

5.2 Speech-based Attack

We describe the modules that constitute *Speech-based Attack*, namely *Pre-processing*, *Training* and *Prediction* modules.

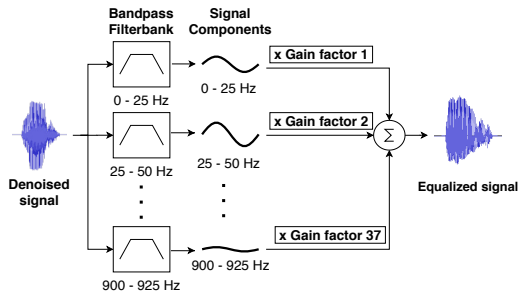

Figure 7: Equalizer implementation using a filterbank of bandpass filters, with associated gain factors for each filter.

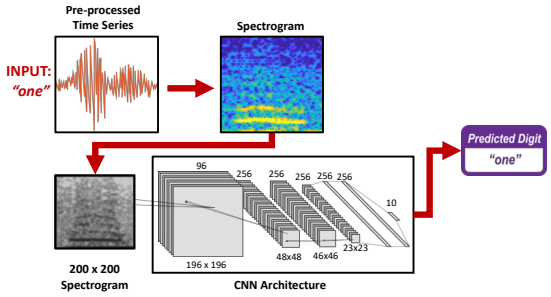

Figure 8: *LidarPhone* CNN architecture, consisting of: one input layer (downscaled spectrogram), four convolutional layers, two dense layers with dropout, and an output layer.

5.2.1 Pre-processing. This module takes as input the signal captured by the lidar sensor and performs a series of pre-processing techniques prior to being input to the subsequent module. The pre-processing step is crucial to the success of the *LidarPhone* attack mainly due to inherently low SNR in the lidar signal. This is attributed to the following reasons. First, the object (such as a trashcan or a takeaway bag) is not in contact with the speaker (i.e., *mechanically decoupled*), attenuating a large portion of the signal. Second, the minutely vibrating object also attenuates parts of the signal, while adding noise. In general, many objects tend to further attenuate high frequency components. Third, capturing such vibration with the lidar sensor at a certain distance away also contributes to additional source of attenuation and noise. Hence, we apply the following series of techniques to ultimately increase the SNR of our input signal.

Correcting DC Offset. The received signal may exhibit a *DC offset* due to minute differences in the receiver sensitivity of lidar sensors. In addition, different objects may reflect the laser signal back to the lidar with different intensities (e.g., glossy plastic vs. wood). Given that the DC offset is the mean amplitude of the signal offset from zero, it may contribute to clipping of the higher amplitude portions, reduced volume, and/or distortion of the sound signals. Hence, we level the DC offset by subtracting the mean of the signal from the original signal.

Outlier Removal and Interpolation. A non-negligible proportion of laser signals that the lidar receives are marked by the lidar as invalid (i.e., *outliers*). This is because some reflected laser signals may be lost as the incident beam reflects off glossy portions of objects, while others are corrupted upon transmission from the lidar due to hardware limitations. We observe that these outliers constitute more than 25% loss of data (at least 1 in every 4 points), further reducing the sampling rate from the original 1.8 kHz (see Section 1). Figure 6(a) depicts this phenomenon as an intense noise band centered at 25% of our sampling rate (465 Hz), muffling the actual utterance by comparison. In order to overcome this problem, we remove such outliers and restore some information by utilizing *cubic spline interpolation*, restoring the signal to the original sampling rate of 1.8 kHz. We specifically use cubic interpolation to accurately model the signal while avoiding *Runge's phenomenon* [56] where higher order interpolations lead to unwanted oscillations contributing to additional noise. These steps remove the noise band and greatly improve the SNR of the signal, as seen in Figure 6(b).

Peak Normalization. Upon interpolation, we perform a *peak normalization* by dividing the value at each point by the maximum signal *peak-to-peak* value, scaling signal amplitudes to the range $[-1, 1]$. This is mainly to control for varying parameters including different sound pressure levels of speech emitted from the speakers, and the distance from the lidar to the target object.

High-pass Filter. We apply a 5th order *high-pass* Butterworth filter to the normalized signal, with a cutoff frequency at 0.5 Hz to filter out low frequency components. This is applied because objects are more affected by low frequency signals (e.g., environmental noise like air-conditioning), contributing in large part to the high noise we observe. In addition, this filter also levels, or *detrends*, the gradual decrease of the lidar signal over time. Such decrease is because the objects are minutely pushed in one direction as the sound signal "deforms" the object, especially if the objects are constituted of thin and light objects (e.g., plastic or paper).

Noise Removal. We further denoise the filtered signal, applying *spectral subtraction* as the following operation: $Y(\omega) = X(\omega) - N(\omega)$, where $X(\omega)$ and $N(\omega)$ are the frequency domain spectra of the input signal and its noise component, respectively. Upon the subtraction, we obtain $Y(\omega)$, namely the resulting signal with noise removed. We estimate $N(\omega)$ by searching in the signal for a segment with the *lowest overall energy* (i.e., ambient noise). The attacker cannot obtain $N(\omega)$ definitively due to the attack's opportunistic nature, and the problem is exacerbated by the noise profile changing over time as the target object deforms. To perform this estimation, we segment $X(\omega)$ into windows of 1024 samples each, and set $N(\omega)$ as the lowest energy window within the previous 30 windows. The effect of noise removal is seen in Figure 6(c) as a reduction in intensity of the noise surrounding the sharp yellow frequency bands of the actual utterance.

Equalization. Figure 7 depicts our *equalization* procedure, where we increase the gain of, or "boost", the lower frequency components of the signal. Objects generally attenuate high-frequency components, concentrating most of the useful information in the reflected signal within the lower frequency components.

Hence, we implement our equalizer based on empirical observations of the average frequency response across different objects (see Section 6.4.1) as we play a known "chirp" signal (i.e., a signal increasing linearly in frequency from 10 Hz to 1 kHz over ten seconds) near them. Our equalizer consists of a filterbank of 37

bandpass filters (5th order Butterworth), which are derived from partitioning the frequency spectrum from 0 Hz to 925 Hz (the approximate Nyquist frequency of our signals) into bins of size 25 Hz each. Each filter has an associated gain factor that is derived ahead of time from our average frequency response across the chosen objects. For example, if most objects reflect a signal with a higher magnitude component in the 50 − 75 Hz band compared to 75 − 100 Hz, the gain factor for the 50 − 75 Hz filter would be set higher than the 75 − 100 Hz filter to amplify the existing information.

To apply the equalizer, the input signal is partitioned into components by the filterbank. Each component has its gain increased by a certain factor, and we sum the resulting signals. The effect of the equalization stages is depicted in Figure 7, where the input time-series signal has its lower frequencies amplified after passing through the equalizer. The output of this stage is used as input for the training and prediction phases of the *Speech-based Attack*, and the correlation stage of the *Sound-based Attack*.

5.2.2 Training and Prediction. The *Training* module takes as input pre-processed lidar signals across multiple training data to ultimately train a classification model to be utilized in the subsequent *Prediction* module during the *Attack Phase*. Furthermore, the attacker would input the corresponding ground truth labels based on specific type of *Speech-based Attack* they are planning to launch, namely *Digit*, *Speaker*, or *Gender Inferences*. For example, when training the model for the *Digit Inference*, the labels would correspond to digits such as *"zero"*, *"one"*, ..., *"ten"*.

We implement *LidarPhone*'s classification modules (i.e., *Training* and *Prediction* modules) with a convolutional neural network (CNN) using the *Keras* [8] and *TensorFlow* [1] machine learning frameworks, with the CNN architecture shown in Figure 8. We design our architecture such that the input to the CNN is a 200×200 spectrogram of our pre-processed signals, where the spectrograms are generated by computing the Short-time Fourier Transform (STFT) of the input signals. Specifically, we leverage the spectrograms as input because transforming the signal to this combined time and frequency domain representation allows for consistently better classification accuracies over a raw time-domain signal. CNNs in particular have been shown to learn more discriminative features from frequency domain representations due to their ability to learn complex features, while preventing overfitting through the use of max-pooling and dropout layers [24]. These advantages consistently allow CNNs to outperform traditional classifiers such as SVM that use hand-crafted features such as Mel-frequency cepstral coefficients (MFCCs) [16, 50], and unsurprisingly, many state-of-the-art audio classification systems use similar approaches [27, 28, 71].

After our input layer, our architecture consists of: two convolutional layers with ReLU activations and max-pooling layers between them, two fully connected dense layers with ReLU activations and a dropout rate of 0.5 each, and a softmaxed output layer to create a probability distribution of the predicted classes. Our architecture is inspired in part by the well-known AlexNet [33] architecture for image recognition, with the aforementioned modifications to tailor it for our spectrogram input and audio classification task. The spectrogram is generated for each signal in Python using *librosa* to generate the STFT with 1025 frequency bins and 128 samples per STFT column since this enabled optimal model performance. The spectrograms are then downscaled for two reasons: a CNN can be trained significantly faster with a smaller image size, and larger input vector sizes may lead to the model learning excessively complex features. The latter is problematic due to our relatively small dataset, which may lead to overfitting and poor accuracy on the test set if the model learns features which are too complex and over-specialized to the training set [79]. This module concludes *LidarPhone*'s *Bootstrapping Phase* allowing the attacker to acquire the classification model to be used in the *Attack Phase*.

During the *Attack Phase*, the attacker utilizes the remotely obtained lidar signal to ultimately infer privacy sensitive speech information. The *Prediction* module takes as input the pre-processed signal, along with the trained classification model, and performs the CNN classification to output the predicted information from speech such as spoken digits for *Digit Inference*.

5.3 Sound-based Attack

We now describe the corresponding modules that constitute the *Sound-based Attack*, namely the *Correlation* module.

5.3.1 Correlation. When launching the *Sound-based Attack*, the attacker uses a captured lidar recording from the victim's robot vacuum cleaner of a small music segment (~10 - 20 seconds) to ultimately infer the news introduction music played through the victim's TV soundbar. The attacker first pre-processes the signal utilizing the aforementioned techniques to increase the SNR, and subsequently inputs the signal to the *Correlation* module. The attacker performs a cross-correlation of the captured and pre-processed signal ($x[n]$) against a previously-prepared pool of original k music signals, i.e., $O = \{o_1[n], o_2[n], \cdots, o_k[n]\}$ from shows on popular news channels. $||x||$ and $||o_i||$ represent the total number of samples in x and the original signal, $o_i \in O$, respectively, where $||x|| \leq ||o_i||$.

The cross correlation "slides" x over each original signal in the pool, o_i. At an offset of t samples, $xCorr(t)$ of the captured signal x and an original sample o_i is defined as $xCorr(x, o_i, t) = \sum_{l=0}^{||o_i||} o_i[l]x[l - t + N]$, where $N = \max(||x||, ||o_i||)$. This sliding approach is necessary since the captured signal can be significantly shorter than the original samples, and the signal could be at any offset within any o_i. This is why a cross-correlation is more effective than the CNN classification technique used in the *Speech-based Attack*. We determine the highest correlation score for o_i as $score_{o_i}$, or the highest score of all the $xCorr(x, o_i, t)$ scores across the sliding windows. Subsequently, we compare the scores across all k music signals to determine predicted music, o_{pred}, to be the one with the maximum $score_{o_i}$ as $o_{pred} = \arg\max_{o_i}(score_{o_i})$.

6 EVALUATION

We now evaluate *LidarPhone* to demonstrate its feasibility.

6.1 Prototype and Experimental Setup

Apparatus. We develop the *LidarPhone* prototype on a Xiaomi Roborock S5 [61] – a popular robot vacuum cleaner that is representative of other robot vacuum cleaners on the market that use lidars for mapping purposes [11–13, 34, 44–46, 53]. We reverse engineer the ARM Cortex-M based firmware of the robot based on a prior attack [19], and gain root access of the system using the

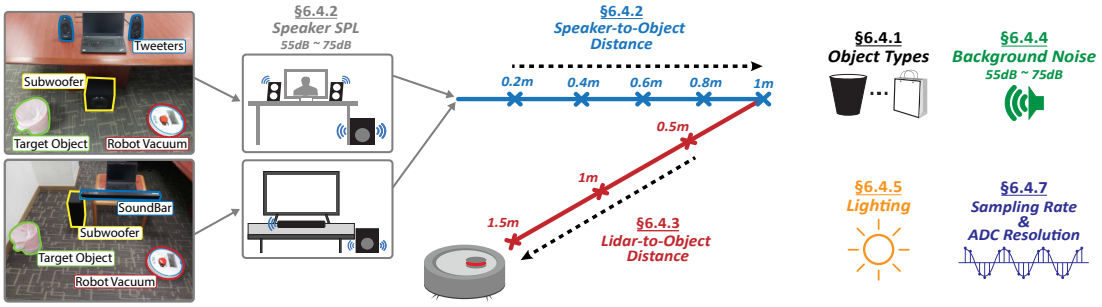

Figure 9: Figure depicts our experimental setup for the *Speech-based Attack* (top left), and *Sound-based Attack* (bottom left). We vary multiple parameters to comprehensively investigate the behavior of *LidarPhone*.

Dustcloud software stack [20]. The robot is typically connected to the Xiaomi cloud ecosystem for its standard operations and data exchange. We override this interface with the *Valetudo* software stack on the rooted device and control the robot over a local network [4].

We design hardware spoofing circuitry to mislead the lidar unit into activating its laser despite not rotating. While a *LidarPhone* attack does not require any hardware modifications, as a proof-of-concept implementation, we implement this step for convenience to avoid changing the lidar's firmware. This gives us access to the raw lidar sensor streams while the robot and the lidar are stationary, and allows us to decode the binary data packets on a remote machine. Hence, we obtain a sampling rate of 1.8 *kHz* from *360 samples per rotation* × *5 Hz rotation frequency* (see Section 1). However, to preserve the robot's ability to navigate using the lidar, we do not interfere with the onboard data processing flow. Rather, we duplicate the lidar data stream on the robot and send it over the wireless network to a laptop using `netcat` [52] for *LidarPhone*'s acoustic processing. The robot then transmits an analog lidar light intensity signal that we process separately offline.

Figure 9 depicts the experimental setups and our evaluation procedure on two realistic home scenarios. For the *Speech-based Attack* we simulate a desktop setup resembling a typical work-from-home (WFH) scenario using a popular desktop speaker set (Logitech Z623 [36]), where the speakers are placed on a desk and the subwoofer is placed on the ground [18]. To simulate a *living room* or *home theater* scenario for the *Sound-based Attack*, we use a common TV soundbar (LG SL5Y [35]). We conduct both experiments in an air-conditioned room to simulate the noise generated by climate control systems in a typical home office or home theater.

Data Collection. We use a portion of the Free Spoken Digit Dataset [25] for *Digit Inference*, consisting of 20 utterances per digit (i.e., 0 to 9). For *Gender Inference* and *Speaker Inference*, we use the TIDIGITS dataset [15], containing speech signals from ten participants (five males and females, respectively), with two utterances per digit per participant for digits "zero" to "nine". Finally, for our *Sound-based Attack*, we construct a dataset of introductory music sequences for ten popular news channels in the U.S. across the conservative-liberal political spectrum [48]. The music sequences are retrieved from YouTube, and consist of five conservative-leaning segments which are coded as FOX [69], FRT [64], FST [68], HAN [65], and SSR [67], and five liberal-leaning segments coded as CNN [43], MER [66], MSN [63], NPR [42] and PBS [62]. In each case, sound is played through the speaker/soundbar near a common household object (see Figure 15 for a full list of *ten objects*) to collect more than 30k utterances totaling over 19 hours of recorded audio.

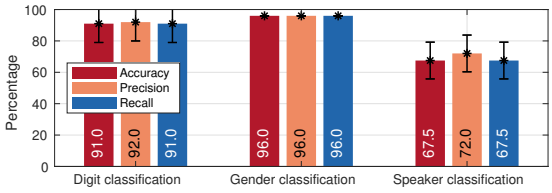

Figure 10: Overall performance of *Speech-based Attack* for different inference tasks.

6.2 Overall Performance

We define a true positive as a captured digit or music segment that is classified correctly by its corresponding classifier, whereas a false negative is a captured segment that is classified as anything other than its correct class. We present the overall performance in terms of accuracy, precision (the ratio of true positives to all positive predictions), and recall (the ratio of true positives to true positives and false negatives) for both *Speech-* and *Sound-based Attacks*.

6.2.1 Speech-based Attack. Figure 10 summarizes the overall performance of the *Speech-based Attack* for different inference tasks. We use a representative and realistic test configuration, where *LidarPhone* targets a *trashcan* (translucent trash bag) that is 20 cm from a speaker emitting 70 dB speech.

		True gender	
		Male	Female
Predicted gender	Male	**96%**	4%
	Female	4%	**96%**

Table 1: Confusion matrix for gender inference.

(i) *Digit Inference*: Figures 10 and 11(a) depict the classification accuracy for ten (0-9) spoken digits for the single speaker case. We achieve a classification accuracy of 91% on average across all digits. This result is significantly higher than a random guess of

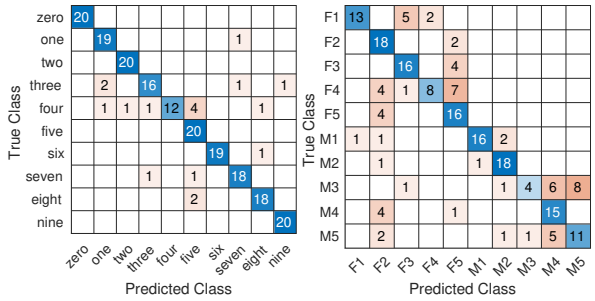

Figure 11: Confusion matrices for (a) digit classification accuracy and (b) speaker classification accuracy, with 20 utterances per class. The *LidarPhone*-processed digit audio and spectrogram files (from "0" to "9") are available for comparison at bit.ly/lidarphone-sensys.

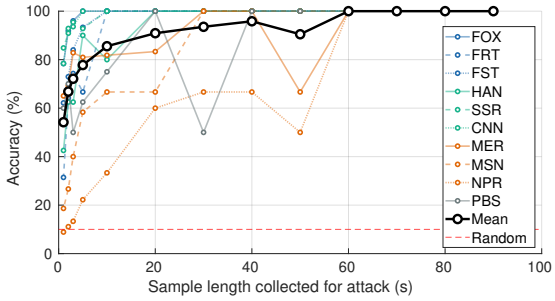

Figure 12: System accuracy for news music classification.

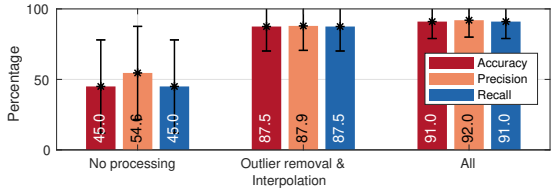

Figure 13: System accuracy with different configurations of our pre-processing pipeline.

10%. The accuracy is primarily impacted by mispredictions of utterances 'three' and 'four'. Specifically, 'four' is mispredicted as 'five' relatively frequently, and further inspection reveals that these two utterances appear almost identical in the spectrogram representations of our recordings. While the original audio spectrograms show distinct features for each digit, the low SNR signals we collect, combined with aliasing effects due to our low sampling rate of 1.8 kHz, reduce their distinguishability in the *LidarPhone* recording. Therefore, the model learns slightly less representative features specific to these digits, and does not classify them as accurately as others.

(ii) *Gender Inference*: We perform gender classification on eavesdropped audio samples to predict the gender of a given speaker, after training on the entire set of male and female speakers. Figure 10 and Table 1 shows a mean accuracy of 96% for this task. We previously hypothesized that male speakers would be classified with higher accuracy than female speakers since higher fundamental voice frequencies, which are more common in females, are likely to fall outside of the frequency band of the recovered sound, leading to higher error. However, the model performs *identically* on male and female cases with 96% classification accuracy. This indicates that our model still manages to capture distinguishing features from high-frequency components that have a relatively low SNR.

(iii) *Speaker Inference*: We present the performance of *LidarPhone* in identifying the current speaker from a set of 10 speakers in Figures 10 and 11(b), where average classification accuracy on this task is 67.5%. We noted that passing only commonly used Mel-frequency spectral coefficient (MFCC) features to our model instead of the raw spectrogram allowed for better performance on this task. Since the MFCCs represent a *reduction* in the available information for the model over the raw spectrogram, we expect that our limited dataset leads to overfitting if the raw spectrogram is used in this specific case. This results in an inability to generalize to new test cases, and therefore worse accuracy. The use of MFCC features improves our accuracy, and we additionally observe that males and females perform comparably, with 64% accuracy and 71% accuracy respectively. This trend matches our observations from our gender classification task.

6.2.2 Sound-based Attack. We evaluate our accuracy at classifying an unknown *LidarPhone* sample of news music as belonging to one of ten possible popular news segments as specified earlier. We use a similar setup to the speech eavesdropping attack for evaluation, namely targeting the same *trashcan* (translucent trash bag) with 70 dB music playing directly adjacent to the object through the soundbar. The length of the collected test sample is varied from 1 to 90 seconds, and the mean accuracy across all classes is plotted alongside accuracy for each individual class in Figure 12. We observe that the system achieves high recognition accuracy (> 90%) with a 20-second sample of music. As the sample length increases, cross-correlation becomes increasingly robust to noise, since there is a lower probability of a long test sample erroneously matching a similar but incorrect signal. This allows the attacker to be flexible; for example, the robot does not have to wait for a time when it can be stationary for a whole minute. The visible fluctuations in accuracy for some classes even as the sample length increases are due to the test set decreasing in size as we combine the samples together. Therefore, a single misprediction has a larger effect on the absolute accuracy value of that class. However, the results are again significantly greater than a random guess of 10%.

6.3 Performance of System Modules

We evaluate the internal modules of the *Digit Inference* attack to compare between alternative designs.

6.3.1 Pre-processing. We evaluate the effect of our pre-processing pipeline (presented in Section 5.2.1) on accuracy. From Figure 13, we can see that outlier removal and interpolation are critical for model accuracy, resulting in an accuracy improvement of 42.5%. We observe from the spectrograms in Figure 6 that the embedded utterance audio in the raw signal has a large noise band caused by outliers. Therefore, removing these outliers and repairing the

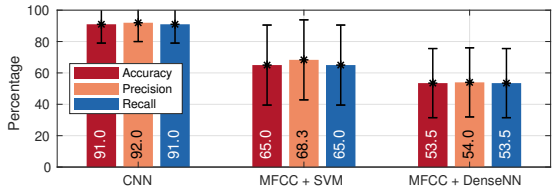

Figure 14: System accuracy with different classifiers.

Figure 15: Different household objects/materials used in *LidarPhone* experiments, in decreasing order of accuracy of digit audio recovery at 70 dB.

missing data points with interpolation has a significant effect. The remaining stages comprise a smaller but non-negligible improvement in accuracy, ultimately increasing it to 91%.

6.3.2 Classification. We evaluate classifiers besides our deep learning model, and show in Figure 14 that our CNN-based classifier significantly outperforms other techniques. Specifically, we evaluate the widely adopted MFCC feature vector with both a traditional Support Vector Machine (SVM) and 4-layer dense neural network architecture. As explained in Section 5.2.2, CNNs are more likely to infer complex but generalizable features from the input spectrogram directly, whereas using the MFCC feature vector in either an SVM or dense network was simply not sufficiently discriminative for our low SNR digit recognition task.

6.4 Impacts of Experimental Conditions

We comprehensively evaluate the performance under different experimental setups and environmental conditions.

6.4.1 Varying Target Objects. Household objects are made of materials having different rigidities and acoustic impedances and therefore respond differently to nearby sounds. We test our digit classification algorithm on the data collected from ten common objects as shown in Figure 15. We select these objects because of the likelihood of finding these on the floor within the reach of the robot's laser. We separate the objects into opaque and matte (mostly diffusely reflective), opaque and glossy (some specularly reflective components), translucent (passing some laser energy through the object), and transparent (passing most of the energy through). We also evaluate collecting signals directly from the subwoofer's front face, which has a metal grill covering a vibrating diaphragm. This attack could be conducted if there are no objects within range of the speaker for *LidarPhone* to target.

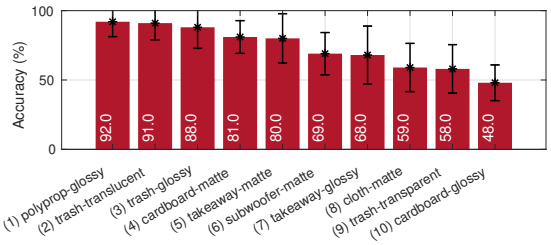

Figure 16: Digit recognition accuracy of all objects at 70 dB speaker volume.

Figure 16 shows that while the expected accuracy for a random guess of the classes is 10%, *LidarPhone* achieves significantly higher scores. Even targeting the subwoofer directly, *LidarPhone* achieves 69% accuracy. Another interesting observation is that relatively more transparent objects (e.g., *cloth* or *trash-transparent*) exhibit insufficient laser reflection, leading to degraded signal quality at the sensor and subsequently lower classification accuracy. We also note that rigidity is an important factor; some objects like *cardboard-glossy* are too flexible, whereas *subwoofer-matte* is too rigid, affecting the SNR of our received signal due to unpredictable or limited vibrations, and leading to lower accuracy scores.

6.4.2 Varying Speaker-to-Object Distance and Speaker Volume. We evaluate our system against changing the distance of the speaker to the object, while also changing the speaker loudness, and present the result in Figure 17(a) for the *trash-translucent* object. We observe a general trend of decreasing system accuracy if the speaker moves further away from the target object or the loudness decreases. This is intuitively expected due to less sound energy incident on the object's surface at further distances and lower volumes.

However, we also observe a rather counter-intuitive effect. When the speaker is close to the target (20 cm), and the speaker's volume is set to its maximum loudness of 75 dB, we observe a *loss* in accuracy compared to the 70 dB case. Upon closer inspection of the recovered spectrograms, we observe that at high effective sound pressure levels (i.e., the combined effect of distance and source sound pressure level), the recovered signal loses a lot of *distinguishing* information, and appears as a uniform *smear* across the frequency domain on the spectrogram. Since all the utterances appear uniform, the model is unable to distinguish useful information to perform classification.

We extend this analysis in Figure 17(b). Here, we once again place the speaker close to the target object at 20 cm, and vary its volume across a representative sample of the objects, with certain interesting properties. As before in Figure 17(a), we notice a loss in accuracy at 75 dB for *trash-translucent*, which is similar for *polyprop*. Interestingly, we notice that for the *takeaway bag*, the loss in accuracy occurs at 70 dB, and for the poorly-performing *subwoofer* object, there is a consistent upwards trend. We believe that this is due to a material-specific "*saturation*" of the target object's surface once a threshold effective loudness is reached. The effect is visible across many of our evaluated objects, and the threshold point also differs across the objects. Note that in Figure 17(a), once the saturation point is reached at the trashcan's level of 75 dB at 20 cm, moving the speaker further away *or* decreasing the source

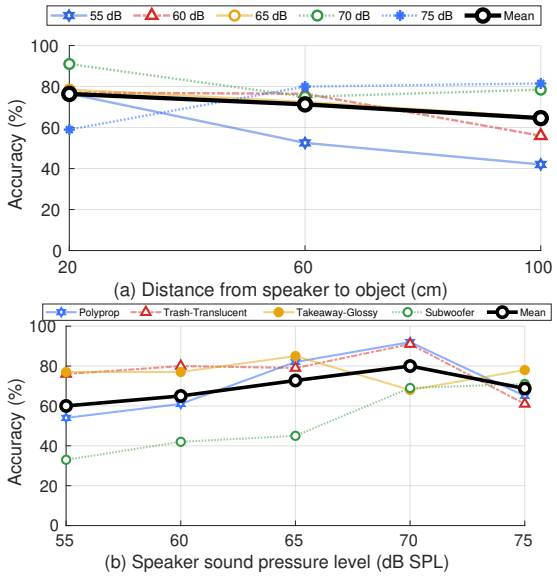

Figure 17: Impact of the source's loudness on the system performance while varying (a) distance from speaker to object and (b) target objects.

volume both improve accuracy. We conclude that for laser-based eavesdropping on different target objects, while louder effective sound volumes improve audio recovery, there is a clear saturation point where no more information can be recovered.

6.4.3 Varying Lidar-to-Object Distance. Figure 18 shows the impact of the distance between the lidar sensor and the object. There appears to be an optimal distance of 150 cm from the lidar to the target where accuracy is optimized. We explain this as a combination of two effects. Light sources *diverge* over longer distances – i.e., if the lidar is further from the object, laser energy is spread onto a larger area on the object surface, and vice-versa. Secondly, object surfaces are not monolithic, and exhibit complex effects on their surface such as resonance and deformation when they vibrate. This can be attributed to the fact that a lidar closer to the target object that focuses laser energy on a smaller area observes less motion of the surface (i.e., lower signal), while also avoiding noise contributed by the non-uniform motion of these surfaces. Therefore, the lidar distance presents a classic signal-to-noise ratio tradeoff, with highest SNR at 150 cm.

6.4.4 Varying Background Noise Levels. We assess the impact of ambient noise levels by playing white noise from the soundbar near the *trash-translucent* object while playing speech for the digit recognition task at 70 dB. Both sound sources are placed equidistant from the target object at 20 cm, and oriented towards the same point on the object. Figure 19 shows the digit recognition accuracy with varying ambient sound levels. Interestingly, the system performance is mostly unaffected even with loud ambient noise as equalization and denoising methods applied during the signal processing steps efficiently eliminate background noise. We only observe significant drops in accuracy at 75 dB and 77 dB.

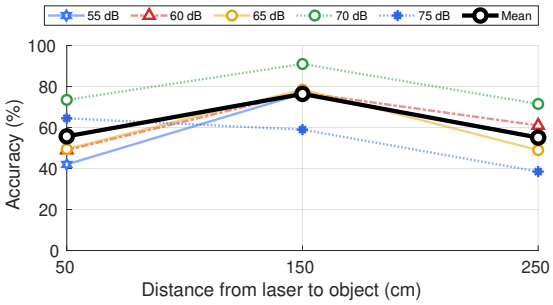

Figure 18: Impact of varying the laser to object distance and speaker volume on digit recognition accuracy.

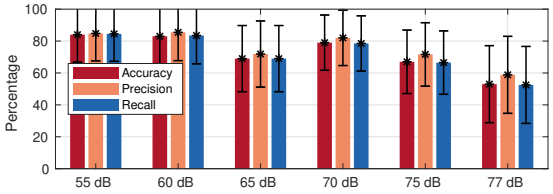

Figure 19: Impact of increasing background white noise levels on digit recognition accuracy.

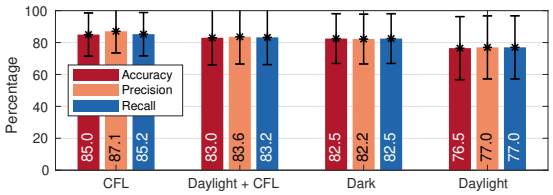

Figure 20: Impact of varying ambient lighting conditions on digit recognition accuracy.

6.4.5 Varying Lighting Conditions. *LidarPhone* depends on the subtle fluctuations of received laser intensities to recover sounds from vibration. Ambient lighting conditions can affect the measured intensity of the optical sensing system in our lidar. Figure 20 shows the performance of the system under different combinations of natural daylight and compact fluorescent lights (CFL). We are robust to differing lighting conditions, except the observation of a small decrease in accuracy in full daylight conditions. The lidar used in our experiments has a laser that is mostly in the infrared spectrum, and the receiver has an infrared filter that blocks all other frequencies of light from reaching the image sensor. Sunlight contains infrared components, which can pollute the received signal, and therefore decrease accuracy. We likely do not see this drop in the Daylight + CFL case since we collect orders of magnitude more training data in this configuration leading to increased model robustness. We also observe that the CFL-only case achieves the highest accuracy, indicating that direct exposure to daylight does present some challenge to *LidarPhone*'s accuracy.

6.4.6 Varying Subwoofer Usage and Placement. It is a common practice to place the subwoofer – the bigger and heavier component of

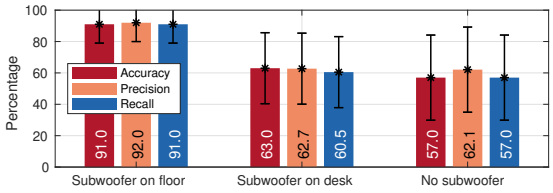

Figure 21: Impact of subwoofer placement on digit recognition accuracy across three cases: (a) Subwoofer placed on the floor (b) Subwoofer placed on the desk (c) No subwoofer.

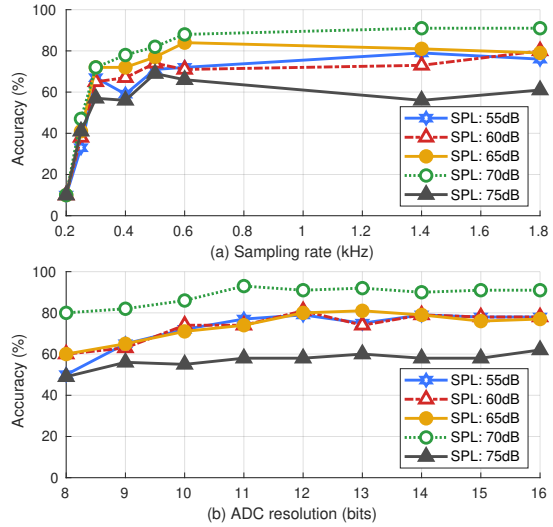

Figure 22: Impact of varying (a) the sampling rate and (b) the ADC resolution on the digit recognition accuracy.

the speaker set – on the floor [18]. However, we also evaluate *LidarPhone*'s performance with various subwoofer placements. Figure 21 compares the average accuracies when the subwoofer is placed on the floor or desk, and when it is not present. With the subwoofer on the desk and the system set to the same volume level, we observe a drop in accuracy. This is expected since the speaker-to-object distance is increased by roughly the height of the table that the subwoofer has been placed on. Without the subwoofer, we once again notice a small drop in accuracy. Since speaker-subwoofer systems are configured such that the subwoofer is responsible for the majority of the power of low frequency components, the loss of the subwoofer leads to significant power loss of those low frequency components and correspondingly, system accuracy.

6.4.7 Varying of Sampling Rate and ADC Resolution. The sampling rate of our lidar sensor and the quantization resolution of the in-built analog-to-digital converter (ADC) affect recorded signal quality. This experiment simulates performing this attack with a lower-quality or cheaper lidar unit. The native sampling rate for the lidar in the Xiaomi Roborock S5 is 1860 Hz, with a 16-bit ADC. Figure 22(a) and 22(b) show the effect of decreasing sampling rate and ADC resolution respectively on the digit recognition accuracy for different sound intensities. From Figure 22(a), the classifier's

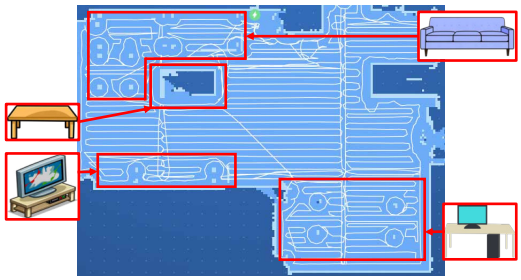

Figure 23: Figure depicts an example of a map generated by a Xiaomi Roborock robot vacuum after a cleaning session. Patterns in the map reveal the locations of common household objects and potential *LidarPhone* attack zones.

performance does not degrade significantly until the sampling rate is below 500 Hz, where fundamental voice frequencies begin to be clipped. From Figure 22(b), only an ADC resolution below 11 bits leads to lower accuracy due to increasing quantization noise. Sensitive equipment like lidars are unlikely to have 11-bit ADCs; most have either 14-bit or 16-bit ADCs, which shows *LidarPhone*'s generalizability to other lidar units. Lastly, the system's dependency on the sound pressure levels remains unchanged for these ADC and sampling rate factors.

7 DISCUSSION

We now discuss the countermeasures, deployment considerations, and limitations of *LidarPhone*.

7.1 Countermeasures

7.1.1 Lidar Rotation. One of the potential defenses against *LidarPhone*'s attack is to further reduce the SNR of the lidar signal. This may be possible if the robot vacuum cleaner lidars are manufactured with a hardware interlock, such that its lasers cannot be transmitted below a certain rotation rate, with no option to override this feature in software. As presented in Section 1, a rotating lidar reduces the sampling rate to 5 Hz for a single point on the target object.

7.1.2 Limiting Side-Channel Information. In our implementation, *LidarPhone* does not use the *distance* reading from the lidar, but instead leverages a "quality" metric that accompanies each reading. This is a noisy but high-resolution value directly related to the *intensity* of the reflected laser beam [58]. We recommend that lidar manufacturers *reduce the resolution* of any user-facing data that directly corresponds to the intensity of reflected laser light.

7.2 Deployment Considerations

Making the Attack More Generic. Recall that our *Digit Inference* (of *Speech-based Attack*) defined in our Threat Model in Section 4 is a *targeted attack*, requiring the adversary to collect the victim's speech for training purposes prior to launching the attack. However, we envision a more *generic attack*, where the adversary may collect a large amount of speech data from multiple individuals to train a generic model, enabling attacks on any desired individual.

While we utilize *digits* as exemplary scenario of capturing privacy sensitive words (including credit card, bank account, and social

security numbers), we may extend *LidarPhone* to capture different privacy sensitive words. Furthermore, we also envision extending *LidarPhone* to discard *unseen* words by utilizing classification confidence scores, where the attacker may be able to launch a more automated approach of capturing privacy sensitive words.

Object Search and Targeting. The adversary may utilize the native mapping features of the robot vacuum cleaner during a *reconnaissance phase* before the actual attack to identify probable target objects. The resulting map exhibits clear patterns as in Figure 23 (e.g., the four legs of a desk), where suitable target objects such as trashcans can be found in close proximity. Afterwards, the robot selects an opportune time for the attack phase, where it targets these locations while seemingly idle.

7.3 Limitations

Recall from Section 6.1 that we prevent the lidar from spinning during a *LidarPhone* attack. While this certainly adds some limitations to the attack, we find that it may be more plausible to launch a *LidarPhone* attack as the robot vacuum cleaner is idle (e.g., docked in its charging station or under furniture). Given that *LidarPhone* is able to launch the attack from a distance away from the object, the victim would still be vulnerable as long as the robot is in the target object's line-of-sight. This approach also renders *LidarPhone* robust from the inherent loud noises generated by its vacuum during cleaning. Additionally, we expect any potential noise caused by individuals walking near the robot vacuum cleaner to be momentary and infrequent disruptions to audio recovery.

Furthermore, we find from our evaluation on subwoofer placement in Section 6.4.6 that *LidarPhone* performs best when the subwoofer is present. Market research data [17] indicates that the trend for possession of home audio systems including subwoofers is increasing, constituting an increasing number of victims vulnerable to *LidarPhone* attacks (assuming that they own lidar-equipped robot vacuum cleaners). However, even without the subwoofer, *LidarPhone* achieves classification accuracy well above a random guess, comparable to other novel side-channel attacks [2, 40]. As an opportunistic attack, *LidarPhone* presents a real threat to victims, as any information gained is beneficial to the attacker.

8 RELATED WORK

We present related work on active vibrometry and passive acoustic eavesdropping.

8.1 Active Vibrometry

Speech signals, like any other sounds, induce vibration in nearby objects and create opportunities for sensing and recovering traces of spoken words by converting this induced vibration to sound. A family of techniques [41, 59, 60, 72] record this vibration by targeting light/LASER beams on the object and measuring the fluctuation of the reflected signal in phase, frequency (doppler shifts [7]), or intensity. A number of recent works [37, 39, 74, 76] demonstrate the measurement of vibration and sound by monitoring changes in reflected wireless radio signals in diverse scenarios including occlusions and non-line-of-sight cases. However, techniques that use LASER vibrometry for acoustic signal recovery are possibly the closest to *LidarPhone*. These techniques optimize hardware for tracking minute variations in received signals to capture vibrations. In contrast, *LidarPhone* shows the possibility of exploiting *existing* lidar sensors on a commodity product for acoustic eavesdropping.

8.2 Passive Acoustic Eavesdropping

Sensors are ubiquitous in our living environments and a range of past techniques have explored their signals as acoustic side-channels. Micro electro-mechanical (MEMS) sensors are used in mobile devices for motion and orientation sensing. Gyrophone [40] is among the first to show that acoustic signals could be captured by MEMS sensors. Later, AccelWord [80] develops techniques to repurpose accelerometers as low-power voice command detectors. Spearphone [2] builds on this core concept to uncover a loophole that bypasses mobile device's permission protocols and records sounds from a smartphone's loudspeaker using its inertial sensors. PitchIn [21] demonstrates the feasibility of speech reconstruction from *multiple* simultaneous instances of non-acoustic sensor (e.g., accelerometer) data collected offline across networked devices. On the other hand, VibraPhone [54] shows that the back-EMF signal of the vibration motor in smartphones and wearables can be processed to recover intelligible speech signals. The Visual Microphone [9] uses high speed video of the target objects to sense vibrations and therefore the nearby source audio. *LidarPhone* faces similar challenges as the Visual Microphone due to indirectly sensing audio through object vibrations. While *LidarPhone*'s core technique and challenges are different from inertial vibration sensing, it shares the same motivation of acoustic side-channels and speech privacy.

9 CONCLUSION

We propose *LidarPhone*, a novel stealthy *remote eavesdropping* attack that exploits the lidar sensor equipped in commodity robot vacuum cleaners, originally used for mapping purposes. *LidarPhone* allows the adversary to obtain privacy sensitive speech information from laser beams reflected off of minutely vibrating objects (such as a trashcan or a takeaway bag) located near the victim's computer speaker or TV soundbar. *LidarPhone* overcomes the limitations of state-of-the-art eavesdropping attacks that require physical presence to deploy eavesdropping equipment, which limits scalability and increases the chances of the attacker getting caught. We implement and evaluate *LidarPhone* to demonstrate its feasibility through real-world experiments. We utilize a vacuum cleaner's lidar sensor to target different objects in its vicinity, collecting speech utterances or music emitted from a computer speaker or TV soundbar, respectively. We demonstrate up to 91% and 90% digit and music classification accuracies, respectively. While we investigate lidars on robot vacuum cleaners as an exemplary case, our findings may be extended to many other active light sensors including smartphone time-of-flight sensors. We hope that this work encourages the SenSys community to investigate appropriate defense mechanisms for such potentially imminent sensor side-channel attacks.

ACKNOWLEDGMENTS

This research was partially supported by a grant from Singapore Ministry of Education Academic Research Fund Tier 1 (R-252-000-A26-133).

REFERENCES

[1] Martín Abadi et al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] S. Abhishek Anand, Chen Wang, Jian Liu, Nitesh Saxena, and Yingying Chen. 2019. Spearphone: A Speech Privacy Exploit via Accelerometer-Sensed Reverberations from Smartphone Loudspeakers. *arXiv:1907.05972 [cs]* (July 2019). http://arxiv.org/abs/1907.05972 00003 arXiv: 1907.05972.

[3] David P Baron. 2006. Persistent media bias. *Journal of Public Economics* 90, 1-2 (2006), 1–36.

[4] Sören Beye. 2020. Hypfer/Valetudo. https://github.com/Hypfer/Valetudo

[5] Steven Boll. 1979. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing* 27, 2 (1979), 113–120.

[6] Bosch. 2019. *ISC-BPR2 - Blue Line Gen2 PIR Motion Detectors.* http://resource.boschsecurity.com/documents/BlueLine_Gen_2_Data_sheet_enUS_2603228171.pdf.

[7] P Castellini, M Martarelli, and EP Tomasini. 2006. Laser Doppler Vibrometry: Development of advanced solutions answering to technology's needs. *Mechanical systems and signal processing* 20, 6 (2006), 1265–1285.

[8] François Chollet et al. 2015. Keras. https://keras.io.

[9] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham Mysore, Fredo Durand, and William T. Freeman. 2014. The Visual Microphone: Passive Recovery of Sound from Video. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 33, 4 (2014), 79:1–79:10.

[10] Dragonmouth and Laura Tucker. 2018. Be Careful of Your Robot Vacuum - It Could Be Spying on You. https://www.maketecheasier.com/robot-vacuum-spying/

[11] ECOVACS. 2020. DEEBOT OZMO 960 - ECOVACS. https://www.ecovacs.com/us/deebot-robotic-vacuum-cleaner/DEEBOT-OZMO-960

[12] ECOVACS. 2020. DEEBOT OZMO T5 - ECOVACS. https://www.ecovacs.com/us/deebot-robotic-vacuum-cleaner/DEEBOT-OZMO-T5

[13] ECOVACS. 2020. DEEBOT OZMO T8 AIVI - ECOVACS. https://www.ecovacs.com/us/deebot-robotic-vacuum-cleaner/DEEBOT-OZMO-T8-AIVI

[14] Paul Eisenstein. 2013. *Seven Best Cars for Front Crash Avoidance.* http://www.thedetroitbureau.com/2013/09/seven-best-cars-for-front-crash-avoidance/.

[15] Dan Ellis. 2003. Clean Digits. https://www.ee.columbia.edu/~dpwe/sounds/tidigits/

[16] Eduardo Fonseca, Rong Gong, Dmitry Bogdanov, Olga Slizovskaia, Emilia Gómez Gutiérrez, and Xavier Serra. 2017. Acoustic scene classification by ensembling gradient boosting machine and convolutional neural networks. In *Virtanen T, Mesaros A, Heittola T, Diment A, Vincent E, Benetos E, Martinez B, editors. Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017); 2017 Nov 16; Munich, Germany. Tampere (Finland): Tampere University of Technology; 2017. p. 37-41.* Tampere University of Technology.

[17] Market Research Future. 2020. Home theatre market research report- forecast 2023 | home theatre industry. https://www.marketresearchfuture.com/reports/home-theatre-market-4121

[18] Dave Gans. 2017. 3 Tips on Where to Place a Subwoofer. https://www.klipsch.com/blog/place-a-subwoofer-3-tips

[19] Dennis Giese. 2018. Having fun with IoT: Reverse Engineering and Hacking of Xiaomi IoT Devices. https://dontvacuum.me/talks/DEFCON26/DEFCON26-Having_fun_with_IoT-Xiaomi.pdf

[20] D Giese. 2018. Having fun with IoT: reverse engineering and hacking of xiaomi IoT devices.

[21] Jun Han, Albert Jin Chung, and Patrick Tague. 2017. Pitchln: eavesdropping via intelligible speech reconstruction using non-acoustic sensor fusion. In *Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks - IPSN '17.* ACM Press, Pittsburgh, Pennsylvania, 181–192. https://doi.org/10.1145/3055031.3055088

[22] Taylor Hatmaker. 2018. A vacuum vulnerability could mean your Roomba knockoff is hoovering up surveillance. https://techcrunch.com/2018/07/19/vacuum-vulnerability-hack-diqee-positive-technologies/

[23] Chris Heinonen. 2015. *Your Privacy, Your Devices, and You.* http://thewirecutter.com/blog/your-privacy-your-devices-and-you/.

[24] Lars Hertel, Huy Phan, and Alfred Mertins. 2016. Comparing Time and Frequency Domain for Audio Event Recognition Using Deep Learning. (2016).

[25] Zohar Jackson, César Souza, Jason Flaks, Yuxin Pan, Hereman Nicolas, and Adhish Thite. 2018. *Jakobovski/free-spoken-digit-dataset: v1.0.8.* https://doi.org/10.5281/zenodo.1342401

[26] Kaspersky. 2018. Xiaomi Mi Robot vacuum cleaner hacked. https://www.kaspersky.com/blog/xiaomi-mi-robot-hacked/20632/

[27] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yong Xu, Wenwu Wang, and Mark D Plumbley. 2019. Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems. *arXiv preprint arXiv:1904.03476* (2019).

[28] Qiuqiang Kong, Turab Iqbal, Yong Xu, Wenwu Wang, and Mark D Plumbley. 2018. DCASE 2018 challenge surrey cross-task convolutional neural network baseline. *arXiv preprint arXiv:1808.00773* (2018).

[29] Kurt Konolige, Joseph Augenbraun, Nick Donaldson, Charles Fiebig, and Pankaj Shah. 2008. A low-cost laser distance sensor. In *2008 IEEE international conference on robotics and automation.* IEEE, 3002–3008.

[30] Kurt Konolige, Joseph Augenbraun, Nick Donaldson, Charles Fiebig, and Pankaj Shah. 2008. A low-cost laser distance sensor. In *2008 IEEE International Conference on Robotics and Automation.* IEEE, Pasadena, CA, USA, 3002–3008. https://doi.org/10.1109/ROBOT.2008.4543666

[31] M Korkmaz, A Durdu, and YE Tusun. 2018. Sensor Comparison for a Real-Time SLAM Application. *International Journal of Information and Electronics Engineering* 8, 1 (2018).

[32] Steven Kreuzer. 2019. *PTPd.* http://ptpd.sourceforge.net/.

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1* (Lake Tahoe, Nevada) *(NIPS'12).* Curran Associates Inc., Red Hook, NY, USA, 1097–1105.

[34] Lenovo. 2020. Lenovo X1: The Newest Lidar-Based Robot Vacuum with Strong Suction and Smart Features. https://smartrobotreviews.com/reviews/lenovo-x1-robot-vacuum-features-review.html

[35] LG. 2020. LG SL5Y : SL5Y 2.1 Channel 400W Sound Bar w/ DTS Virtual: X & High Resolution Audio. https://www.lg.com/us/home-audio/lg-SL5Y

[36] Logitech. 2020. Logitech Z623 2.1 Speaker System with Subwoofer. https://www.logitech.com/en-us/product/speaker-system-z623

[37] Yongsen Ma, Gang Zhou, and Shuangquan Wang. 2019. WiFi sensing with channel state information: A survey. *ACM Computing Surveys (CSUR)* 52, 3 (2019), 1–36.

[38] Meghan McDonough. 2018. Roborock s5 robot vacuum review: jack-of-all-trades, master of none. https://www.tomsguide.com/us/roborock-s5-robot-vacuum,review-6274.html

[39] William McGrath. 2005. Technique and device for through-the-wall audio surveillance. US Patent App. 11/095,122.

[40] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. 2014. Gyrophone: Recognizing Speech from Gyroscope Signals. In *23rd USENIX Security Symposium (USENIX Security 14).* USENIX Association, San Diego, CA, 1053–1067.

[41] Ralph P Muscatell. 1984. Laser microphone. US Patent 4,479,265.

[42] Television Music. 2016. NPR Morning Edition Theme Song. https://www.youtube.com/watch?v=otGmpQpiVjw&list=PLtxqRD7TEIMdPsuCVV1fK30zQzdJ9gBtn&index=10&t=0s

[43] Television Music. 2018. The Situation Room Theme Music - CNN. https://www.youtube.com/watch?v=IpoXtx6mkPY&list=PLtxqRD7TEIMdPsuCVV1fK30zQzdJ9gBtn&index=2&t=0s

[44] Neato Robotics. 2020. Botvac™ Connected | Neato Robotics | Singapore |. http://www.neatorobotics.com.sg/botvac-connected/

[45] Neato Robotics. 2020. Neato D4 robot vacuum – Neato – Intelligent Robot Vacuums. https://neatorobotics.com/products/neato-d4/

[46] Neato Robotics. 2020. Neato D6 robot vacuum – Neato – Intelligent Robot Vacuums. https://neatorobotics.com/products/neato-d6/

[47] O'Donnell. 2019. IoT Robot Vacuum Vulnerabilities Let Hackers Spy on Victims. https://threatpost.com/iot-robot-vacuum-vulnerabilities-let-hackers-spy-on-victims/134179/

[48] University of Michigan. 2020. Where do news sources fall on the political bias spectrum? https://guides.lib.umich.edu/c.php?g=637508&p=4462444

[49] Theodor Olsson and Albin Larsson Forsberg. 2019. IoT Offensive Security Penetration Testing. (2019).

[50] Sangwook Park, Seongkyu Mun, Younglo Lee, and Hanseok Ko. 2017. Acoustic scene classification based on convolutional neural network using double image features. In *Proc. of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017).* 98–102.

[51] Pery Pearson. 1993. *Sound Sampling.* http://www.hitl.washington.edu/projects/knowledge_base/virtual-worlds/EVE/I.B.3.a.SoundSampling.html.

[52] Avian Research. 2007. Netcat: the TCP/IP swiss army. https://nc110.sourceforge.io/

[53] Neato Robotics. 2020. Neato D7 robot vacuum – Neato – Intelligent Robot Vacuums. https://neatorobotics.com/products/neato-d7/

[54] Nirupam Roy and Romit Roy Choudhury. 2016. Listening through a Vibration Motor. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services - MobiSys '16.* ACM Press, Singapore, Singapore, 57–69. https://doi.org/10.1145/2906388.2906415

[55] RpLidar. 2014. Low cost 360 degree 2D laser scanner (Lidar) system–Introduction and Datasheet. *Robopeak Team* (2014).

[56] Carl Runge. 1901. Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten. *Zeitschrift für Mathematik und Physik* 46, 224-243 (1901), 20.

[57] Argo-A Security. 2020. Long-Range Laser Listening Device. http://argoasecurity.com/index.php?route=product/product&product_id=263

[58] SLAMTEC. 2018. RPLIDAR Interface Protocol and Application Notes. http://bucket.download.slamtec.com/b42b54878a603e13c76a0a0500b53595846614c6/LR001_SLAMTEC_rplidar_protocol_v1.1_en.pdf

[59] G Smeets. 1977. Laser interference microphone for ultrasonics and nonlinear acoustics. *The Journal of the Acoustical Society of America* 61, 3 (1977), 872–875.

[60] John R Speciale. 2001. Pulsed laser microphone. US Patent 6,301,034.

[61] Roborock Technology. 2020. Roborock S5 Max Robot Vacuum & Mop Cleaner. https://us.roborock.com/pages/roborock-s5-max

[62] Television Music. 2015. Music by David Cebert PBS News Hour Open Theme. https://www.youtube.com/watch?v=gzofWrHsmK4&list=PLtxqRD7TEIMdPsuCVV1fK30zQzdJ9gBtn&index=8&t=0s

[63] Television Music. 2018. The 11th Hour Theme Music - MSNBC. https://www.youtube.com/watch?v=44b5L-vn3_4&list=PLtxqRD7TEIMdPsuCVV1fK30zQzdJ9gBtn&index=3&t=0s

[64] Television Music. 2018. Fox Report Theme Music - Fox News. https://www.youtube.com/watch?v=ov8lAenbJUg&list=PLtxqRD7TEIMdPsuCVV1fK30zQzdJ9gBtn&index=9&t=0s

[65] Television Music. 2018. Hannity Theme Music - Fox News. https://www.youtube.com/watch?v=xyzO0kA0ZX0&list=PLtxqRD7TEIMdPsuCVV1fK30zQzdJ9gBtn&index=4&t=0s

[66] Television Music. 2018. MSNBC Election Result Theme (Loop). https://www.youtube.com/watch?v=vlnKoDzCZxE&list=PLtxqRD7TEIMdPsuCVV1fK30zQzdJ9gBtn&index=7&t=0s

[67] Television Music. 2018. Shepard Smith Reporting Theme Music - Fox News. https://www.youtube.com/watch?v=Khr8wCB-G0U&list=PLtxqRD7TEIMdPsuCVV1fK30zQzdJ9gBtn&index=13&t=0s

[68] Television Music. 2018. The Story Theme Music - Fox News. https://www.youtube.com/watch?v=xak5VCT4_do&list=PLtxqRD7TEIMdPsuCVV1fK30zQzdJ9gBtn&index=11&t=0s

[69] Television Music. 2020. Fox News Democracy 2020 Theme Music. https://www.youtube.com/watch?v=aWk9DypnyPI&list=PLtxqRD7TEIMdPsuCVV1fK30zQzdJ9gBtn&index=5&t=0s

[70] AV Test. 2019. From the Land of Smiles – Xiaomi Roborock S55. https://www.iot-tests.org/2019/02/from-the-land-of-smiles-xiaomi-roborock-s55/

[71] Michele Valenti, Aleksandr Diment, Giambattista Parascandolo, Stefano Squartini, and Tuomas Virtanen. 2016. DCASE 2016 acoustic scene classification using convolutional neural networks. In *Proc. Workshop Detection Classif. Acoust. Scenes Events*. 95–99.

[72] Chen-Chia Wang, Sudhir Trivedi, Feng Jin, V Swaminathan, Ponciano Rodriguez, and Narasimha S Prasad. 2009. High sensitivity pulsed laser vibrometer and its application as a laser microphone. *Applied Physics Letters* 94, 5 (2009), 051112.

[73] Winston Wang. 2019. Calculating Political Bias and Fighting Partisanship with AI. https://www.thebipartisanpress.com/politics/calculating-political-bias-and-fighting-partisanship-with-ai/

[74] Teng Wei, Shu Wang, Anfu Zhou, and Xinyu Zhang. 2015. Acoustic eavesdropping through wireless vibrometry. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. 130–141.

[75] Lawrence B. Wolff, Diego A. Socolinsky, and Christopher K. Eveland. 2003. Using infrared sensor technology for face recognition and human identification. In *Infrared Technology and Applications XXIX*, Bjorn F. Andresen and Gabor F. Fulop (Eds.), Vol. 5074. International Society for Optics and Photonics, SPIE, 757 – 766. https://doi.org/10.1117/12.498156

[76] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, Kun Wang, and Wenyao Xu. 2019. Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 14–26.

[77] YDLIDAR. 2020. YDLIDAR G6. https://www.ydlidar.com/service_support/download.html?gid=6

[78] Yida. 2020. What is a Time of Flight Sensor and How does a ToF Sensor work? https://www.seeedstudio.com/blog/2020/01/08/what-is-a-time-of-flight-sensor-and-how-does-a-tof-sensor-work/

[79] Xue Ying. 2019. An overview of Overfitting and its solutions. In *Journal of Physics: Conference Series*, Vol. 1168. IOP Publishing, 022022.

[80] Li Zhang, Parth H. Pathak, Muchen Wu, Yixin Zhao, and Prasant Mohapatra. 2015. AccelWord: Energy Efficient Hotword Detection Through Accelerometer *(ACM MobiSys)*. https://doi.org/10.1145/2742647.2742658