PrivacyMic: Utilizing Inaudible Frequencies for Privacy Preserving Daily Activity Recognition

Yasha Iravantchi
University of Michigan
Ann Arbor, MI
yiravan@umich.edu

Karan Ahuja
Carnegie Mellon University
Pittsburgh, PA
kahuja@cs.cmu.edu

Mayank Goel
Carnegie Mellon University
Pittsburgh, PA
mayank@cs.cmu.edu

Chris Harrison
Carnegie Mellon University
Pittsburgh, PA
chris.harrison@cs.cmu.edu

Alanson Sample
University of Michigan
Ann Arbor, MI
apsample@umich.edu

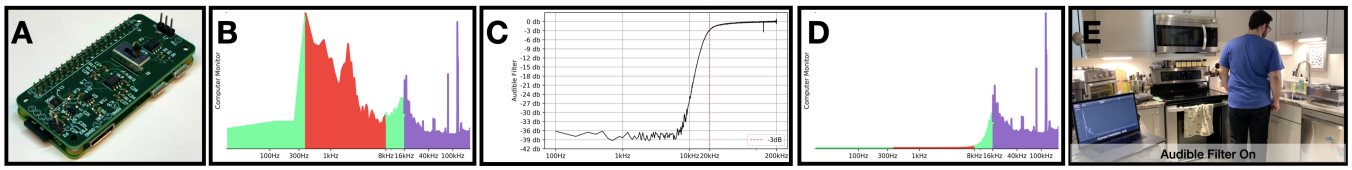

Figure 1: (A) PrivacyMic's Pi Zero-based hardware (B) captures audible (green) and ultrasonic (purple) frequencies and (C) uses in-hardware filters to (D) remove privacy-sensitive speech (red) frequencies to (E) perform privacy-preserving activity recognition.

ABSTRACT

Sound presents an invaluable signal source that enables computing systems to perform daily activity recognition. However, microphones are optimized for human speech and hearing ranges: capturing private content, such as speech, while omitting useful, inaudible information that can aid in acoustic recognition tasks. We simulated acoustic recognition tasks using sounds from 127 everyday household/workplace objects, finding that inaudible frequencies can act as a substitute for privacy-sensitive frequencies. To take advantage of these inaudible frequencies, we designed a Raspberry Pi-based device that captures inaudible acoustic frequencies with settings that can remove speech or all audible frequencies entirely. We conducted a perception study, where participants "eavesdropped" on PrivacyMic's filtered audio and found that none of our participants could transcribe speech. Finally, PrivacyMic's real-world activity recognition performance is comparable to our simulated results, with over 95% classification accuracy across all environments, suggesting immediate viability in performing privacy-preserving daily activity recognition.

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445169

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; *Ubiquitous and mobile computing*; Ubiquitous and mobile computing systems and tools.

KEYWORDS

Acoustics, Ultrasound, Infrasound, Internet of Things, IoT, Microphones, Smart Environments, Sound Sensing, Ubiquitous Sensing

ACM Reference Format:
Yasha Iravantchi, Karan Ahuja, Mayank Goel, Chris Harrison, and Alanson Sample. 2021. PrivacyMic: Utilizing Inaudible Frequencies for Privacy Preserving Daily Activity Recognition. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3411764.3445169

1 INTRODUCTION

Microphones are perhaps the most ubiquitous sensor in computing devices today. Beyond facilitating audio capture and replay for applications such as phone calls and connecting people, these sensors allow computers to perform tasks as our digital assistants. With the rise of voice agents, embodied in smartphones, smartwatches, and smart speakers, computing devices use these sensors to transform themselves into listening devices and interact with us naturally through language. Their ubiquity has led them to find other purposes beyond speech, powering novel interaction methods such as in-air and on-body gestural inputs [39, 45]. More importantly, microphones have found use within health sensing applications, such as measuring lung function and performing cough detection [16, 23]. While the potential of ubiquitous IoT devices is limitless, the ever-present, ever listening microphone presents significant

privacy concerns to users. This conflict leaves us at a crossroads: How do we capture sounds to power these helpful, always-on applications without capturing intimate, sensitive conversations? The current "all-or-nothing" model of disabling microphones in return for privacy throws away all the microphone-based applications of the past three decades.

Typically, the microphones that drive our modern interfaces are primarily designed to operate within human hearing — roughly 20Hz to 20kHz. This focus on the audible spectrum is perhaps not surprising given these microphones are most often used to capture sounds for transmission or playback to other people. However, removing the speech portion of the audible range reduces the accuracy of audible-only sound classification systems, as speech makes up almost half of the audible range. Fortunately, there exists a wealth of information beyond human hearing: in both infrasound and ultrasound. The human-audible biases in sound capture needlessly limit computers' ability to utilize sound. However, useful, inaudible acoustic frequencies can be used to generate new sound models and perform activity recognition, entirely without the use of human-audible sound. Furthermore, these inaudible frequencies can replace privacy-sensitive frequency bands, such as speech, and compensate for the loss of information when speech frequencies are removed.

In this work, we seek to explore sounds outside of human hearing and their utility for sound-driven event and activity recognition. As an exploration into inaudible sounds, we built a wide-band audio capture apparatus capable of recording infrasonic, audible, and ultrasonic frequencies. We used this apparatus to collect audio from 127 commonplace items (e.g., faucet, gas furnace, microwave, light bulbs) across three homes and four commercial buildings. We use this dataset to answer two primary questions: (1) Do our daily-use objects emit significant infrasonic and ultrasonic sounds? (2) If the devices do emit these sounds, are these inaudible frequencies useful for recognition? From our spectral analysis, we found all of our objects emit inaudible frequencies that contain significant information power: almost 43% exists outside of human hearing.

Our inaudible sounds evaluation thus informed our design of PrivacyMic, a wide-range microphone with in-hardware filters that can remove speech frequencies or all audible frequencies entirely. To facilitate easy deployment in a wide variety of applications, PrivacyMic is fully Raspberry Pi compatible, shares the same footprint as a Raspberry Pi Zero, and draws less than 1mA, allowing for mobile and battery-powered applications. We then used PrivacyMic to perform a user study simulating eavesdropping where participants listened in on filtered speech: none of our participants could transcribe a single word in the audio clip on either filter setting. These results suggest that the captured audio is unintelligible to the naked ear and provides a degree of privacy for always-on listening devices. Finally, we performed a real-world study, performing activity recognition tasks using speech-filtered or audible-filtered input across three common environments, kitchen, bathroom, and office, and found over 95% accuracy across all environments and filter settings. By performing privacy-preserving activity recognition, we hope this work provides an alternative to the "all-or-nothing" model and inspires greater adoption of inaudible frequencies as privacy-preserving signal sources for acoustic activity recognition.

2 RELATED WORK

While there are many direct sensing approaches that instrument specific objects for detecting human activity, such as water-pressure sensors [13], powerline sensors [17], and sensor fusion approaches [24], we focus more closely on acoustic-only methods to contextualize our work. We primarily look at sound classification and labeling systems, which try to predict the sound in an audio clip (e.g., dog barking, doorbell ringing). Building upon these systems are acoustic activity recognition systems, which, similar to PrivacyMic, work on the assumption that sound events have a direct inference on human activity (e.g., if the sound detected is a microwave, a person must be using it). We also look at other work that more broadly uses inaudible sounds as sensing mechanisms. Finally, we review works that explore privacy-related concerns with microphones.

2.1 Collected Sound Databases and Labeling

Sound labeling systems focus on predicting the content of an audio segment. Classical machine learning approaches employ Fast Fourier Transforms (FFTs) and Mel Frequency Cepstral Coefficients (MFCCs) as features and use conventional supervised learning algorithms for classification. For example, Foggia et al. [11] use a bag of words model with Support Vector Machines (SVMs) as a classifier. Hidden Markov Models (HMMs) have also been employed for context recognition as showcased by Eronen et al. [10]. In more recent work, deep learning has shown exceptional performance in sound labeling tasks. These approaches rely on convolutional neural networks (CNNs) for classification, omitting the need to create hand-crafted features. For example, AudioSet [14] can label all the events that occurred within a sound clip. NELS [9] presents a never-ending learner of sounds, a system that continuously learns from the web relations between sound and language, and uses a CNN for event detection. Sound has also been used for scene and object recognition, as presented in SoundNet [4]. Works by Cakir et al. [6] have also exploited the temporal nature of sound and have used convolutional recurrent neural networks to create sound models. Most of these approaches use thousands of labeled data points in the audible spectrum and rely heavily on tightly curated and well-labeled sound databases to develop and train their models. As a result, these models cannot be easily extended to utilize inaudible frequencies.

Supporting these models are sound effect libraries, which provide an abundance of labeled sound data. These sound effects are pure and atomic, meaning that the clip's sound is tightly segmented and contains only a single class of sound. Additionally, within the last 20 years, compiled and crowdsourced-labeled datasets, such as FreeSound [12] and AudioSet [14], have become increasingly available. These sound sources are not necessarily atomic and pure but do provide variety to improve machine learning methods' generalizability. Unfortunately, both of these datasets have varying sampling rates from 16kHz to 48kHz. While a 48kHz recording may, in theory, extend into ultrasound (i.e., Nyquist limit = 24kHz), it is uncertain that the microphone used for recording had a frequency

response in either the infrasonic or ultrasonic range. Thus, the collected sound databases available today are unsuitable for training an inaudible sound-based activity recognition models.

2.2 Acoustic Activity Recognition

Acoustic activity recognition systems build upon sound labeling works by using a sound classification model to infer activity. These works rely on the assumption that a sound event is a direct result of human activity. The most prevalent example in commercial applications is ShotSpotter [40], which performs audible gunshot detection. Synthetic Sensors [24] uses a sensor fusion approach, but relies most heavily on microphone data for activity recognition. In terms of more generalized sound-based activity recognition models, Ubicoustics [23] performs acoustic activity recognition using standard audible microphones in laptops and mobile devices. These systems purposefully downsample audio input to 16kHz (Nyquist limit = 8kHz) to reduce computational overhead and training set homogeneity, as audio datasets contain audio clips of varying sampling rates. This downsampling causes the resultant audio to omit useful higher frequencies and rely heavily on speech frequencies (300Hz $< f <$ 8kHz) as input. PrivacyMic takes advantage of this omitted source of information and showcases the utility of inaudible bands for privacy-preserving sensing.

2.3 Infrasound and Ultrasound for Sensing

While infrasound has been used to evoke human sensory responses, such as reproducing the effects of earthquakes and interactive vibrating sculptures, it has a relatively small body of work in HCI applications as a sensing method. Kijima et al. built a system detecting door opening and closing [22], and Chang et al. used infrasound sensors to determine the state of windmill farms [7]. Ultrasound has a much greater body of previous work, such as being used for hand [19] and facial [20] gesture detection, including imaging within the body [28]. Further, ultrasound has been used as a tool for indoor localization [38] and data transfer [15]. Despite an extensive literature search, we could not find prior work that passively collects infrasound and ultrasound and treats them as extensions of audible sound as a sensing method.

2.4 Privacy and Microphones

As the number of microphone-containing devices increases in our daily lives, there is significant interest in works that look to both exploit privacy issues and assuage privacy concerns [27, 31]. Systems that look to attack microphones can inject audio to initiate unauthorized commands to voice agents using lasers [41], ultrasonic side channels [2], non-linearities in commodity smartphone microphones [36]. From a privacy-preserving sensing perspective, Larson et al. explore cough detection by creating a carefully crafted feature set that can reproduce cough sounds but not speech [25]. Zhang et al. explore privacy-preserving Parkinson's disease tracking by discarding all but non-speech body sounds [46]. Lee et al. explore algorithms that garble speech, but keep other biologically-relevant sounds intact for lung disease tracking [26]. However, while the final features obscure speech and ensure a degree of privacy, these works perform the filtering in software, allowing for the possibility that unfiltered audio can leave the audio capture device. PrivacyMic performs speech and audible filtering in hardware before the ADC, such that 1) it is robust to voice injection attacks since speech frequencies are aggressively filtered and 2) an attacker cannot modify the software of PrivacyMic to allow privacy-sensitive audio to leave the device. We discuss these hardware filters further in our Hardware Implementation section and in our Privacy Evaluation.

Figure 2: An image of the wide-band audio capture rig, consisting of infrasound, audible, and ultrasonic microphones, along with a camera. The set-up is used to capture sounds from 127 different objects and devices commonly found in residential homes and commercial buildings. The data is used to determine which frequency components are most predictive of object and device usage.

3 INAUDIBLE SOUNDS COLLECTION

Given the number of animals that can hear sub-Hz infrasound (e.g., whales, elephants, and rhinos) [43] and well into ultrasound (e.g., dogs to 44kHz, cats to 77kHz, dolphins to 150kHz) [43], it is perhaps unsurprising that there is a world of exciting sounds around us that we cannot hear. While these animals have adapted their hearing to meet their evolutionary needs, such as for long-distance communication, hunting prey, and echolocation, human hearing, and subsequently microphones, have been tuned to capture human sounds and speech [34]. We designed an information power study to explore the inaudible world and answer two fundamental questions: (1) Do our daily-use objects emit significant infrasonic and ultrasonic sounds? (2) If the devices do emit these sounds, are these inaudible frequencies useful for recognition?

3.1 Wideband Capture Apparatus

To collect sounds from three distinct regions of the acoustic spectrum, we built an audio-capture rig (see Figure 2) that combines three microphones with targeted frequency responses: infrasound, audible, and ultrasound. While these microphones have overlap in frequency responses, we define acoustic frequency ranges and source signals from the appropriate microphone with the least attenuation to create a "hybrid" microphone. We describe our methodology in detail in further sections. The microphones are all connected via USB to a standard configuration 2013 MacBook Pro 15" for synchronized data capture. The internal microphone in the MacBook Pro was also captured as an additional audible source for possible future uses. We also added a webcam to provide video recordings of the objects in operation. FFMpeg (Fast Forward Moving Pictures Expert Group) was used to simultaneously capture from all audio sources and the webcam, synchronously. FFMpeg was configured

to use a lossless WAV codec for each of the audio sources (set to the appropriate sampling rate) and H.264 with a QScale of 1 (Highest Quality) for the video recording. These choices were to ensure that no losses due to compression occurred in the data collection stage.

We define infrasound as frequencies below human hearing (f < 20Hz) [30]. To capture infrasonic acoustic energy, we used an Infiltec INFRA20 Infrasound Monitor, via a Serial-to-USB connector. The INFRA20 has a 50Hz sampling rate with a pass-band from 0.05Hz to 20Hz [18]. While the sensor itself has a frequency response above 20Hz (Figure 3, top), the device has an analog 8 Pole elliptic filter with a 20Hz corner frequency low pass filter. As a result, we do not use the INFRA20 to source acoustic signal for any other acoustic region.

While humans can detect sounds in a frequency range from 20Hz to 20kHz, this is often in ideal situations and childhood, whereas the upper limit in average adults is often closer to 15-17kHz [33]. We quickly confirmed this by having adult colleagues listen to 15-17kHz test tones and all had their hearing limits within this range. For this paper's purposes, we define the upper limit of audible as the midpoint of that range (i.e., what can be expected for an average adult to be able to hear), resulting in a total audible range of 20Hz < f < 16kHz. To capture audible signals, we used a Blue Yeti Microphone set to Cardioid mode to direct sensitivity towards the forward direction with a gain of 50%. [5]. The Yeti has a 48kHz sampling rate and a measured frequency response of 20Hz to 20kHz (Figure 3, middle) [35]. While our ultrasonic microphone's frequency response includes the Yeti's entirely, the Yeti had less attenuation from 10kHz to 16kHz. As a result, we source our audible signal solely from the Yeti.

For ultrasound frequencies (f > 16kHz), we used a Dodotronic Ultramic384k [8]. The Ultramic384k has a 384kHz sampling rate, with a stated frequency range up to 192kHz. The Ultramic384k uses a Knowles FG-series Electret capsule microphone. While we could not find a response curve from Dodotronic directly, we found the following response curve from Knowles for the FG microphone from 10kHz up to 110kHz (Figure 3, bottom) [3]. In laboratory testing, we found the Ultramic384k continues to be responsive above 110kHz to the Nyquist limit of 192kHz and as low as 20Hz. Our Ultramic384k had less attenuation than the Yeti from 16kHz to 20kHz (the upper limit of the Yeti), resulting in our ultrasound signal sourced solely from the Ultramic384k.

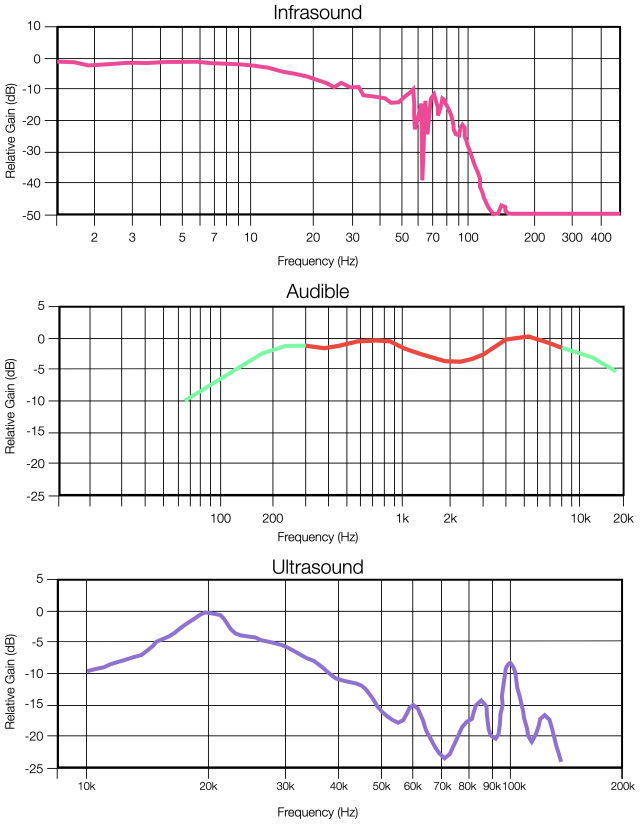

Figure 3: Individual frequency response curves for the infrasound, audible, ultrasonic microphones plotted on log-log axis. The optimal frequency response for each microphone is used to synthetically construct a single microphone from 0.05Hz to 192kHz.

3.2 Data Collection Procedure

To introduce real-world variety and many different objects, including different models of the same item (e.g., Shark vacuum vs. Dyson vacuum), data was collected across three homes and four commercial buildings. More information about these locations and a full list of all these objects can be seen in Table 1. In the real world, sensing devices are not always afforded the luxury of perfectly direct and close sensing. We assessed, after some preliminary testing of how far and off-angle we can place the rig while capturing good signal, that a 45° angle at a distance of 3 m are reasonable parameters (less than -12 dB attenuation across all microphones) to simulate conditions experienced by a sensing device in the home or office while still retaining good signal quality. It is important to note that this is not the distance limit of PrivacyMic, see Figure 7. For some of our items, physical constraints (e.g., small spaces like kitchens and bathrooms) prevented us from measuring at those angles and distances. In those cases, a best effort was made to maintain distances and angles that would be expected in a real-world sensor deployment.

Before recording the object, a 5-second snapshot was taken as a background recording to be used later for background subtraction. Almost immediately after, the item was activated, and a 30-second recording was performed. Five instances of background recording and item recording were captured for each item. For items that do not require human input to continue operation, such as a faucet, the item was turned on prior to the beginning of the 30-second recording, but after the 5-second snapshot, and left on for the entirety of the clip. For an item that required human input, such as flushing a toilet, the item was repeatedly activated for the entire duration of the clip (i.e., every toilet clip has multiple flushes). The laptop's microphone and video from the webcam on the rig were also captured in the clips for potential future use. If multiple items were being recorded in the same session, we would rotate through items in a random order so that none of the objects' sounds were

collected back-to-back to avoid having their clips sound too similar or be contiguous. If only one item was being captured in that session, the rig would be moved and re-placed prior to recording. This was to prevent the capture from being identical and adds variety for machine learning classification and would result in accuracy numbers more indicative of real-world performance. Lastly, if objects had multiple "modes" (e.g., faucet normal vs. faucet spray), we captured both as separate instances.

3.2.1 Homes. We collected sounds in 3 homes: one apartment, one townhome, and one single-family single-story home. We focused primarily on recording kitchen and bathroom sounds, but also captured additional interesting objects. 71 of our 127 sounds were sourced in homes. In the kitchen, we focused primarily on capturing sounds from kitchen appliances such as blenders and coffee makers as well as commonly found fixtures such as faucets and drawers. Overall, 30 different kitchen objects were collected across three homes. In the bathroom, we focused mainly on capturing sounds from water-based sources such as toilets and showers. Additionally, we captured everyday grooming objects, such as electric toothbrushes, electric shavers, and hairdryers. Overall, 24 different bathroom objects were collected across three homes. Apart from those two contexts, we captured general home items, such as laundry washers and dryers, vacuum cleaners, and shredders. We also collected audio from two vehicles, one motorcycle and one car. This resulted in an additional 17 objects collected across two of the three homes.

3.2.2 Commercial Buildings. We also wanted to collect sounds in commercial buildings, as the general nature of similar objects differs and introduces a variety of different objects. We chose four different environments across four commercial buildings: workshops, office spaces, bathrooms, and kitchenettes. We also collected sounds from objects of interest that did not fit in those four categories. 56 of our 127 sounds were sourced in commercial buildings. The workshop contained primarily power tools such as saws and drills, as well as specialized tools, such as laser cutters and CNC machines. We also captured sounds from fixtures such as faucets and paper towel dispensers. Overall, 12 objects were sourced from one of the four commercial buildings. The commercial bathroom, similar to the home bathroom, focused on water-based sounds from toilets and faucets but also contained sounds from things not commonly found in home bathrooms like paper towel dispensers and stall doors. This environment contributed 16 objects from three of the four commercial buildings.

The kitchenette consisted of small office/workplace-style kitchens containing microwaves, coffee machines, and sometimes dishwashers and faucets. This environment contributed to 18 objects from two of the four commercial buildings. The office space contained sounds such as doors, elevators, printers, and projectors, contributing 6 distinct sounds from one of the four commercial buildings. The miscellaneous category contained sounds that were collected in the commercial buildings but did not fit in the above four categories. This included items such as vacuums and a speaker amplifier, contributing 4 items from one of the four commercial buildings.

4 INAUDIBLE SOUNDS EVALUATION

To evaluate the importance of each region of acoustic energy, we first featurize our raw signals using a log-binned Fast Fourier Transform (FFT), which we then analyze using information power metrics. Finally, we use these metrics to perform classification tasks using different combinations of features sourced from distinct acoustic regions. We want to explicitly state that this study is meant to strictly evaluate the utility of different acoustic spectra and is not an evaluation of real-world performance (which we evaluate in our Real World Evaluation). We now describe each section in detail.

4.1 Featurization

In order to provide features for feature ranking and machine learning, we first create a high-resolution FFT of the infrasound, audible, and ultrasound recordings, for both the background and the object. We then perform background subtraction, subtracting the background FFT components from the object's FFT. This allows us to create a very clean FFT signature of solely the object, which minimizes the machine learning models from learning the background rather than the object itself. Since the 127 classes were collected across different locations, the background information aids by effectively segmenting the classes by location, which helps separate similar objects (e.g., sinks vs sinks) due to their different backgrounds.

However, using fixed bin sizes with 0.1Hz resolution, the resulting feature vector contains approximately 2 million features. Therefore, to maintain high frequency resolution at low frequencies while keeping the number of features reasonable, we composite a 100 log-binned feature vector from 0Hz to 192kHz. This resulted in 27 infrasound bins, 53 audible bins, and 20 ultrasound bins. These feature vectors (and subsets of these vectors) will be used as inputs both for our Feature Ranking tasks and our classification tasks. The feature bins can be seen in Figure 4.

While it is prevalent for sound-based methods to use Mel-frequency cepstral coefficients (MFCCs), we opt for FFTs due to their versatility in capturing the signal outside of human-centric speech. MFCCs are widely used for speech recognition and employ the Mel filter bank, which is optimized for human hearing and auditory perception [44]. As humans are better at discerning pitch changes at low frequencies rather than higher ones, the Mel filter bank becomes broader and less concerned with variations for higher frequencies [44]. Therefore, while great for detecting human speech, which has a fundamental frequency from 300Hz and a maximum frequency of 8kHz [42], it allocates a large portion of the coefficients in that low fundamental frequency range and performs poorly in capturing the discriminative features at higher frequency ranges as its resolution decreases.

4.2 Spectral Information Power

To quantify the importance of each spectral band, we employ feature selection methods that rank each band by its information power. There are several ways this can be done, including unsupervised feature selection or dimensionality reduction methods, such as Principal Component Analysis (PCA). However, since we have a well-labeled dataset, we can perform supervised feature selection and classification using Random Forests, which are robust and can

Table 1: Comprehensive table of the 127 objects and devices recorded with the wide-band audio capture rig, along with location and environmental information such as room type and size.

Location	Dims (m) [L]	[W]	[H]	Objects	Wall Materials	Notes
Apartment: Kitchen/Living Room	14.4	6.8	2.4	Air Filter, Blender, CFL Bulb, Dishwasher, Kitchen Drawer, Faucet (Normal), Faucet (Spray), Humidifier, Ice Maker, Computer Keyboard, KitchenAid Mixer, Microwave, Printer, Refrigerator Door, Paper Shredder, Anova Sous Vide, Toaster Oven, Shark Vacuum, Stove Vent, Water Boiler	Drywall/Glass	Room contains windows
Apartment: Bathroom	4.6	2.4	2.4	Bathroom Drawer, Clothes Dryer, Face Scrubber, Faucet, Hair Dryer, Electric Shaver, Shower (Normal), Shower (Spray), Toilet Flush, Electric Toothbrush, Clothes Washer, Waterpik	Drywall/Glass	Room contains glass mirror and glass shower walls
Apartment: Car Garage	15.0	6.0	3.5	Car	Concrete	Apartment complex indoor garage parking structure
Townhome: Kitchen	13.2	4.5	2.6	Blender, Coffee Grinder, Faucet (Normal), Faucet (Spray), Milk Frother, Garbage Disposal, Microwave, Refrigerator Door, Refrigerator Water Dispenser, SodaStream, Stove, Stove Vent, Toaster, Toaster Oven, Water Boiler	Drywall/Wood/Glass	Room contains windows
Townhome: Workshop	15.1	3.5	2.2	Horizontal Saw, Tool Box	Stone	Room is in basement, has building supports
Townhome: Bathroom	5.9	4.2	2.4	Bathtub Faucet, Floor Heater, Hair Dryer	Drywall	Room contains windows
Townhome: Laundry Room	3.8	1.5	2.4	Clothes Dryer, Hot Water Heater, HVAC Furnace	Drywall	
Townhome: Hallway	3.5	1.5	2.4	Fireplace	Drywall	Room contains windows, hallway that connects to other rooms
Townhome: Outdoors	N/A	N/A	N/A	Motorcycle	N/A	Outdoors open area
Home: Kitchen	3.2	3.2	2.4	Dishwasher, Milk Frother	Drywall/Wood/Glass	Room contains windows
Home: Bathroom	3.6	2.8	2.4	Bath Faucet, Bathroom Cabinet, Bathroom Drawer, Ceiling Fan, Sink Faucet, Electric Shaver, Shower, Toilet Flush, Electric Toilet Seat, Electric Toothbrush, Waterpik	Drywall/Glass	Room contains glass mirror and glass shower walls
Commercial1: Bathroom	5.6	2.5	2.4	Faucet, Paper Towel Dispenser, Soap Dispenser, Stall Door, Toilet Flush, Urinal Flush	Drywall + Metal	Bathroom stalls are metal
Commercial1: Workshop	13.4	5.0	12.7	CNC Mill, Air Compressor, Drill Press, Dust Gorilla, Faucet, Angle Grinder, Hand Drill, Laser Cutter Vacuum, Paper Towel Dispenser, Rotary Saw, Shop Vac, Table Saw	Drywall	Room contains machine shop equipment
Commercial1: Lab Space	13.8	5.5	2.6	Coffee Machine, Door Close, Dyson Vacuum, Elevator, Espresso Machine, Hair Dryer, Hand Vacuum, Ice Maker, Microwave, Printer, Projector, Refrigerator Door, Sliding Door, Speaker Amp, Toaster Oven, Water Fountain	Drywall/Metal/Glass	Conjoined with Commercial1: Kitchenette
Commercial1: Kitchenette	9.2	5.5	2.6	Blender, Cabinet, Coffee Grinder, Dishwasher, Faucet (Normal), Faucet (Spray), Microwave, Refrigerator Door, Water Boiler	Drywall/Wood/Glass	Conjoined with Commercial1: Lab Space
Commercial1: Storage Hallway	4.4	1.2	2.6	Paper Towel Dispenser, Faucet	Drywall/Wood	Conjoined with Commercial1: Kitchenette
Commercial2: Kitchenette	4.8	2.2	3.2	Faucet, Microwave, Paper Towel Dispenser	Drywall/Wood	
Commercial3: Bathroom	5.2	2.5	2.4	Faucet, Paper Towel Dispenser, Toilet Flush, Urinal Flush	Drywall + Metal	Bathroom stalls are metal
Commercial4: Bathroom	6.4	3.2	2.4	Faucet, Paper Towel Dispenser, Toilet Flush, Urinal Flush	Drywall + Plastic	Bathroom stalls are plastic

build a model using the Gini impurity-based metric [29]. Using the Gini impurity to measure the quality of our split criterion, we can quantify the decrease in the weighted impurity of the feature in the tree, which indicates its importance [37]. Another critical aspect of Random Forests is that it decreases the importance of features already duplicated by other features: given a spectral band that has high importance and another spectral band that represents a subset of the same information, the importance of the latter will be reduced. As our goal is not to study the relationship between features but to quantify the singular importance of each band, this metric allows us to quantify the standalone information power of each band.

We found that half of the top 10 features were in the ultrasound range, the remainder in the audible range. Figure 4 (top) shows the top 20 feature importances sorted from greatest to least. Of the top 20 features, all audible features are within the privacy-sensitive speech range. Figure 4 (bottom) shows the feature importance sorted by frequency. In both figures, infrasound is denoted in pink, audible in green, and ultrasound in purple. Further examination shows that for infrasound, features below 1Hz have 0 information power. We suspect this is because we could not capture a significant number of objects that emit sub-Hz acoustic energy and only two of our objects (HVAC Furnace and Fireplace) had the majority of their spectral power in infrasound. We found that below 210Hz, there is a gradual tapering of feature importance for audible frequencies, which we suspect is due to a similar reason. For ultrasound, our greatest components came in the low ultrasound region (f<50kHz), which also contained 5 of our top 10 components. The average importance for infrasound, audible, and ultrasound was 0.006, 0.011, and 0.013. Infrasound (27 bins), audible (53 bins), and ultrasound (20 bins) contributed 16.2%, 57.8%, and 26% of the total information power, respectively.

4.3 Classification Accuracy

We quantify our results of spectral analysis in terms of classification accuracies as well. We test our system across various frequency ranges, devices, and a privacy-preserving mode, described below. For our evaluation, we use a Random Forest Classifier with 1000 estimators and evaluate performance in a leave one round out cross-validation setting. Given that we have five instances of each class type, we divided the training set into four instances of each class, and the corresponding test set contains one instance of each class, across five rounds. Other techniques such as Support Vector Machines and Multi-Layer Perceptron achieved similar performances; hence we decided to remain with a Random Forrest Classifier to make our classification machine learning pipeline match our spectral information power.

We wanted to quantify the usefulness of each frequency band in terms of its impact on activity recognition. Therefore, we passed the frequency bins of each spectrum and evaluated its performance. When using Infrasound alone (f < 25Hz), our system achieves a mean classification accuracy of 35.0%. For Audible alone (20Hz <

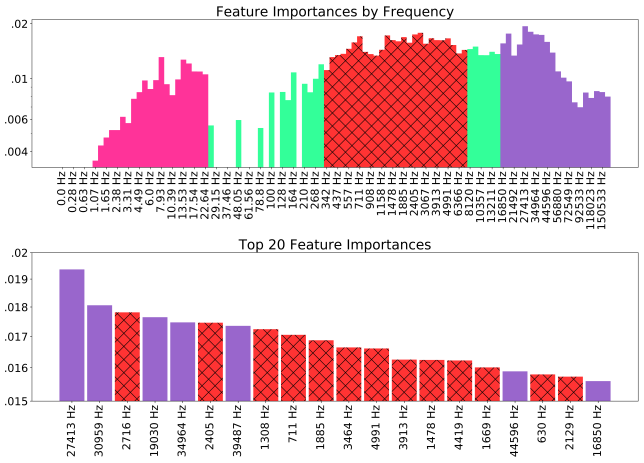

Figure 4: The bar plot on the top shows each frequency component's predictive power, as measured by Gini Impurity. Infrasound is presented in pink, audible frequencies in green, and ultrasonic components in purple. Bins that contain privacy-sensitive speech components are denoted with crosshatches, indicating frequencies that should be avoided. The bar plot on the bottom shows the top 20 most important frequencies ranked in order of their Gini Impurity. This shows that ultrasound is the most viable frequency range for creating a privacy-preserving always-on microphone.

Table 2: The classification accuracies per different combinations of frequency ranges and their privacy-preserving status. Classification performance is restored to >90% when adding Ultrasound to speech removed Audible.

Input Frequencies	Accuracy	Privacy Preserving
Infrasound	35.0%	**Yes**
Audible - Speech	50.5%	**Yes**
Ultrasound	70.2%	**Yes**
Infrasound + Ultrasound	80.2%	**Yes**
Audible	89.9%	No
Audible + Ultrasound - Speech	90.3%	**Yes**
Full Spectrum - Speech	91.4%	**Yes**
Audible + Ultrasound	92.8%	No
Audible + Infrasound	93.2%	No
Full Spectrum	95.6%	No

Infrasound: f < 25Hz, Speech: 300Hz < f < 8kHz, Audible: 20Hz < f < 16kHz, Ultrasound: f > 16kHz

f < 16kHz), we get an accuracy of 89.9%. For Ultrasound alone (f > 16kHz), we get an accuracy of 70.2%. When using the Full Spectrum of acoustic information (i.e., all ranges combined), we found a mean classification accuracy of 95.6%. We also performed pairwise combinations of the three regions; their results can be seen in Table 2. We observe an accuracy increase of 5.7% when using Full Spectrum over using Audible alone.

It is interesting to note that compact fluorescent lightbulbs (CFLs) and humidifiers have powerful ultrasonic components, with minimal audible components, and are only distinguishable in that band. The fireplace has more significant components in infrasound than in ultrasound and audible, and the HVAC furnace solely emits infrasound. The mutual information from all bands also helps to build a more robust model for fine-grained classification. Particularly interesting are items that sound similar to humans, such as water fountains and faucets, which are confused in audible ranges, but can be distinguished when using ultrasonic bands. Also, items such as projector and toaster oven, which were misclassified by each band individually, were only correctly predicted when combining all frequency bands' information.

We investigate the use of our system in a privacy-preserving mode as well. That is, the performance of our system when no speech components, from 300Hz to 8000Hz to include higher-order harmonics, can be captured by it. To simulate this setting, we drop the speech frequencies from our features. Therefore, we test three conditions: audible frequency ranges without speech, audible and ultrasound without speech, and full-spectrum without speech. We found a significant drop in performance when removing speech frequencies from audible, from 89.9% to 50.5%. We find that when using privacy-preserving audible + ultrasound and full-spectrum, our algorithm retained robustness, suffering an accuracy drop of only 5.3% and 4.2%, respectively. Table 2 shows the results of removing speech frequencies relative to other frequency combinations and their privacy-preserving status.

5 HARDWARE IMPLEMENTATION

From our findings in the information power study, we set out to design a microphone optimized for high-audible and ultrasonic frequencies, and avoiding frequency ranges that contain potentially private information. We omit using an infrasound microphone for PrivacyMic as they are physically large. Additionally, we only encountered a few number of objects that were infrasound-heavy (e.g., HVAC Furnace, Fireplace), providing only a 1.1% improvement in classification performance. PrivacyMic consists of three major components: a wide-band ultrasonic microphone, in-hardware amplification and filter stage, and a low-noise, low-power high speed Analog to Digital Converter (ADC). We now describe those components in greater detail.

5.1 System Architecture

In order to faithfully capture high-audible and ultrasonic frequencies, we needed to select a microphone that had sufficient range (8kHz-192kHz) and could be filtered in-hardware. In-hardware filtering removes privacy-sensitive frequencies, such as speech, in an immutable way, preventing an attacker from gaining access to sensitive content remotely or by changing PrivacyMic's software. In-hardware filtering also ensures that no speech content will ever leave the device when set to speech or audible filtered, since the filtering is performed prior to the ADC. PrivacyMic does not have access to these frequency components. While there are a number of Pulse Density Modulation (PDM) microphones that would fulfill our frequency range requirements, performing in-hardware filtering is significantly easier in the analog domain. Thus, since we had extensive experience with the Knowles FG microphone in the Dodomic, we use this microphone as our input. Since the Knowles FG microphone produces small signals ($25mV_{pp}$), we preamplify these signals with an adjustable gain (default G = 10) prior to filtering. The preamp is connected to a double pole triple throw switch, connecting the amplified signal to a high pass speech filter (f_c = 8kHz), an audible filter (f_c = 16kHz), or directly passed through unfiltered. We will describe these filters in further detail in our

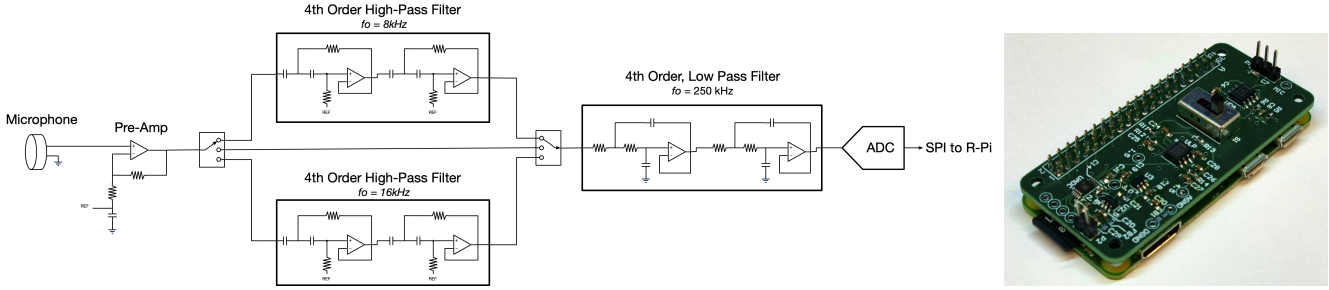

Figure 5: Block diagram of the PrivacyMic daughter board shown on the left and an image of the fully assembled system connected to a Raspberry Pi Zero W on the right. The PrivacyMic hardware uses an analog ultrasonic microphone and employees a 4th order high pass Sallen-Key filter to remove privacy invasive speech and/or audible frequencies. A 250kHz low pass filter is required by the 500kHz, SAR Analog to Digital Converter, to remove spurious high frequency interference. This prototype uses switchable signal paths to test different cut-off frequencies.

hardware evaluation section. Past the filters, the signals are then passed to a low-pass filter set to the Nyquist limit of the ADC (f_c = 250kHz) to remove aliasing, high frequency noise, and interfernce. Finally, a high-speed low-power SAR ADC samples these signals (up to 500kHz) and is connected to a Raspberry Pi Zero W via SPI. The Pi Zero then performs the SPI transaction and sends each sample to a computer via TCP. Figure 5 shows the full schematic of PrivacyMic.

5.2 Hardware Evaluation

In this section, we describe in further detail the performance of our in-hardware speech and audible filters, how well PrivacyMic can pickup sounds from a distance, and the power consumption and wireless performance of PrivacyMic.

5.2.1 Filter Performance. We wanted to evaluate the performance of our speech and audible filters. Instead of performing a frequency sweeps using a speaker and microphone, which introduces inconsistencies through the frequency response of the microphone and output speaker, we bypassed the microphone and provided input directly to our filters using a function generator [21]. For both filters, we also performed a linear sweep and a log sweep from 100Hz to 100kHz and found significant signal suppression below the filter cutoff. Figure 6 shows the the filter filter performance of our speech filter (top) and our audible filter (bottom). While our measured filters' cutoffs are more aggressive than our 8kHz and 16kHz targets, we still found reasonable classification performance (see Real World Evaluation).

5.2.2 Distance Performance. To evaluate how well our microphone is able to pickup sounds from a distance, we selected an audible speaker [1] and a piezo transducer [32] and drove the speaker/transducer at different frequencies using a function generator [21], set the output to high impedance and amplitude to $10V_{pp}$. While the impedances of the speakers were not equal, we do not make comparisons across or between speakers. In order to minimize the effects of constructive and destructive interference due to reflections, we selected a large, empty room (18m long, 8.5m wide, 3.5m tall) to perform our acoustic propagation experiments. We marked distances of 1m, 2m, 4m, 6m, 9m, 12m, and 15m at an angle of 0° (direct facing), placing the microphone at each distance resulting in 7 measurements per frequency. For each measurement, we calculated the RMS for the given test frequency (i.e., the signal was filtered and all other frequency components/noise removed). We then normalized the values of each distance to the max RMS value for that frequency. We fit an exponential curve in the form $y = a * e^{-b*x} + c$ fit to the data. Figure 7 shows that across multiple frequencies, our microphone is able to pick up signal well above the noise floor even 15m away. It is important to note that while PrivacyMic does not use any frequencies below 8kHz, they were included for comparative purposes.

5.2.3 Power Consumption. We measured the power consumption of PrivacyMic using an ammeter. At 5V, we found PrivacyMic draws .6mA and the Raspberry Pi Zero W draws 180mA during operation. A 1600mAh battery, which has the same footprint as the Pi Zero, could sustain continuous operation for 7 hours.

5.2.4 Wireless Performance. We found the average latency to track with the latency of the network, with an average overhead of 2ms on top of the network latency (i.e., ping times of 8ms resulted in

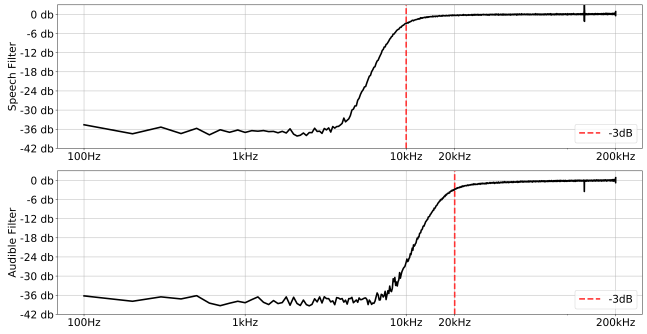

Figure 6: The magnitude Bode plots generated from linear sweeps of our speech and audible filters from 100Hz to 200kHz.

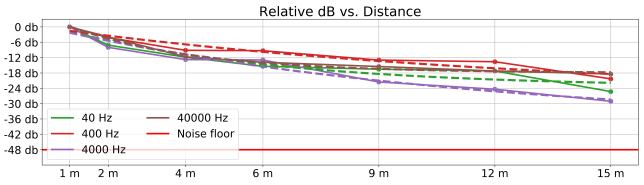

Figure 7: The distance response curves across four test frequencies, with the noise floor denoted in red.

overall latency of 10ms). We were also able to saturate the wireless bandwidth of the Raspberry Pi Zero W in testing, but only require up to 16Mbps at a 500kHz sampling rate (500,000 samples * 4 bytes/sample * 8 bits/byte).

6 PRIVACY EVALUATION

There are numerous privacy concerns surrounding always-on microphones in our homes placed in locations where they have access to private conversation. Two possible avenues where microphones can be compromised are bad actors gain access to audio streams off the device directly or through mishandled data breaches. We performed a user study, evaluating whether our participants were able to perceive various levels of content within a series of audio clips, as if they were an eavesdropper listening to a PrivacyMic audio stream. We also used this evaluation to confirm our previously selected frequency cutoffs of 8kHz for speech and 16kHz for audible.

6.1 Procedure

We generated 3 audio files by reading a selected passage from Wikipedia for approximately 30 seconds using PrivacyMic. For file A, we used our speech filter, removing all frequencies below 8kHz. We noticed that while speech frequencies were removed, some higher frequency fragments of speech remained in the speech filtered file. To simulate a potential attack vector, we pitch shifted the harmonic frequencies down to 300Hz (the lower range of human voice frequencies), and generated file B. For file C, we used our audible filter, removing all frequencies below 16kHz. All of the files were saved as a 16-bit lossless WAV. We recruited 8 participants (Table 3) and asked them, for each of the files, to respond on a Likert scale (1 to 7, 1 being "Not at all" and 7 being "Very clearly") to the questions seen in Table 3.

We also elicited general comments per file and comments comparing the three files. The partcipants were asked to wear headphones for this study, but we did not restrict volume, they were permitted to increase or decrease volume to their preference and listen to the clip multiple times.

6.2 Results

For file A, which had all speech frequencies removed, had mixed responses on whether the participants could hear *something* in the file. However, participants were in general agreement that they could not hear human sounds and were almost unanimous that they could not hear speech. The ones that said they could hear speech stated "someone speaking but not intelligable" and "it sounds like grasshoppers but the cadence of the sounds seems like human speech". All participants agreed with a score of 1 that they could not hear speech well enough to transcribe. None were able to transcribe a single word from the audio clip.

For file B, which was the pitch shifted version of file A, more participants stated that they could hear something in the file, and a greater number stated that they were human sounds, but again the majority could not identify the sound as speech: "it sounded like someone was breathing heavily into the mic" and "it sounds like a creepy monster cicada chirping and breathing". All but one participant stated with a score of 1 that they could not hear speech well enough to transcribe. None were able to transcribe a single word from the audio clip.

For file C, which had all audible frequencies removed, had fewer participants than file A or B report that they could hear things in the file. Additionally, all but one reported with a score of 1 that they could attribute the sounds to human, and all but one reported with a score of 1 that they were able to hear speech. The same participant who recognized the cadence in file A also reported "Sounds like tinny, squished mosquito. Could make out the cadence of human speech". None were able to transcribe a single word from the audio clip.

Overall, only one participant was able to identify the content as speech in two of the files, but none were able to transcribe a single word from any of the files.

6.3 NLP Results

Additionally, we processed our audio files through various natural language processing services (CMU Sphinx, Google Speech Recognition, Google Cloud Speech to Text) and found none of them were able to detect speech content within the files. All of these services were able to transcribe the original, unfiltered audio correctly.

While we do not discount that there may be avenues to reconstruct speech from these speech and audible filtered clips, these results show promise in preventing a bad actor from discerning conversations by "listening in" alone.

7 REAL WORLD PERFORMANCE

While our Inaudible Sounds Evaluation presents promising results, we also evaluated the performance of PrivacyMic in a less controlled environment. Rather than consistently placing the microphone 3m and 45° from the object, the microphone is placed in a natural location relative to its environment in this real-world evaluation, which introduces variety and realism. Unlike the previous study, we do not perform background subtraction, and the objects remain in their natural setting, allowing for a mixture of real-world noise, volumes, and distances.

7.1 Procedure

We placed PrivacyMic near an electrical outlet for each environment, similar to typical IoT sensor placement such as an Alexa. We collected ten rounds for each object in that environment, capturing ten instances per round, 3000 samples per instance. Since we do not evaluate across environments in this evaluation (and real-world systems do not have the luxury of background subtraction), we

Table 3: Questions asked of participants per each file, summary statistics, and demographic information.

Question	Mean	SD
For file A: Were you able to hear anything in the file?	4.5	2.1
For file A: Were you able to hear human sounds in the file? (i.e., sounds that could be coming from a human)	2.3	2.1
For file A: Were you able to hear speech in the file?	1.5	1.1
For file A: Were you able to discern speech well enough to transcribe?	1.0	0.0
General comments on Sound File A (what did you hear if at all, what you perceived was in the file, can you transcribe)	N/A	N/A
For file B: Were you able to hear anything in the file?	4.9	1.8
For file B: Were you able to hear human sounds in the file? (i.e., sounds that could be coming from a human)	2.6	1.2
For file B: Were you able to hear speech in the file?	1.4	0.7
For file B: Were you able to discern speech well enough to transcribe?	1.3	0.7
General comments on Sound File B (what did you hear if at all, what you perceived was in the file, can you transcribe)	N/A	N/A
For file C: Were you able to hear anything in the file?	3.1	1.8
For file C: Were you able to hear human sounds in the file? (i.e., sounds that could be coming from a human)	1.3	0.7
For file C: Were you able to hear speech in the file?	1.1	0.4
For file C: Were you able to discern speech well enough to transcribe?	1.0	0.0
General comments on Sound File C (what did you hear if at all, what you perceived was in the file, can you transcribe)	N/A	N/A
General comments on each sound file (what did you hear if at all, what you perceived was in the file) compared to one another	N/A	NA
What is your age?	25.9	2.3
What gender do you identify as?	M: 5, F: 2, NB: 1	N/A
What level of education are you at?	Grad: 8	N/A

Figure 8: PrivacyMic was evaluated in three real-world environments: (A) kitchen, (B) bathroom, (C) office.

do not collect a background clip for background subtraction. Additionally, for each environment, ten rounds of the "nothing" class were also collected, where none of the selected objects were on. We also do not control for items turning on in the background (such as a refrigerator or A/C). This procedure was repeated for both the speech filter and the audible filter.

We performed our real-world evaluation in 3 familiar environments similar to our previous evaluation: kitchen, bathroom, and office. These environments can be seen in Figure 8 and in the Video Figure. For the kitchen environment, we selected the kitchen sink, the microwave, and a handheld mixer. For the office environment, we selected writing with a pencil, using a paper shredder, and turning on a monitor. For the bathroom environment, we selected an electric toothbrush, flushing a toilet, and the bathroom sink.

7.2 Results

After collecting the data, we performed a leave-one-round-out evaluation, where we trained on nine rounds and tested on the tenth, all combinations results averaged. We featurize the waveforms and use a Random Forest classifier (1000 estimators) in a similar pipeline to Section 4. However, as stated earlier, this evaluation omits background subtraction and also must correctly predict "nothing" when no object is in use.

7.2.1 Speech Filter Results. We found performance consistent with our earlier results using the speech filter, where frequencies less than 16kHz are removed. For the kitchen environment, we found an average accuracy of 99.3% (SD = 1.1%). For the bathroom environment, we found an average accuracy of 99.7% (SD = 0.8%). For the office environment, we found an average accuracy of 99.3% (SD = 1.1%). We also explored the performance of a unified model, where we performed a leave-one-round-out evaluation on all 10 classes. In order to prevent a class imbalance (as there are three times the number of instances for the nothing class), we perform us the nothing class from each environment separately and average the results. For our unified model, we found an average accuracy of 98.9% (SD = 0.7%). The confusion matrices for each condition can be found in Figure 10.

7.2.2 Audible Filter Results. We found performance consistent with our earlier results using the audible filter, but slightly degraded compared to the speech filter, where frequencies less than 16kHz are removed. For the kitchen environment, we found an average

accuracy of 95.0% (SD = 2.7%). For the bathroom environment, we found an average accuracy of 98.2% (SD = 2.2%). For the office environment, we found an average accuracy of 99.3% (SD = 1.6%). Similar to the speech filter results, we evaluated the performance of a unified model, and found an average accuracy of 95.8% (SD = 2.1%). The confusion matrices for each condition can be found in Figure 10.

8 DISCUSSION

While classification accuracies suggest that the audible range is the most critical standalone acoustic range, the average importance of each bin was greater in ultrasound by 18% compared to audible, making it the most valuable region per bin. When restricting input frequencies to only "safe" frequency bands, classification accuracies suggest a different story: ultrasound alone provides an almost

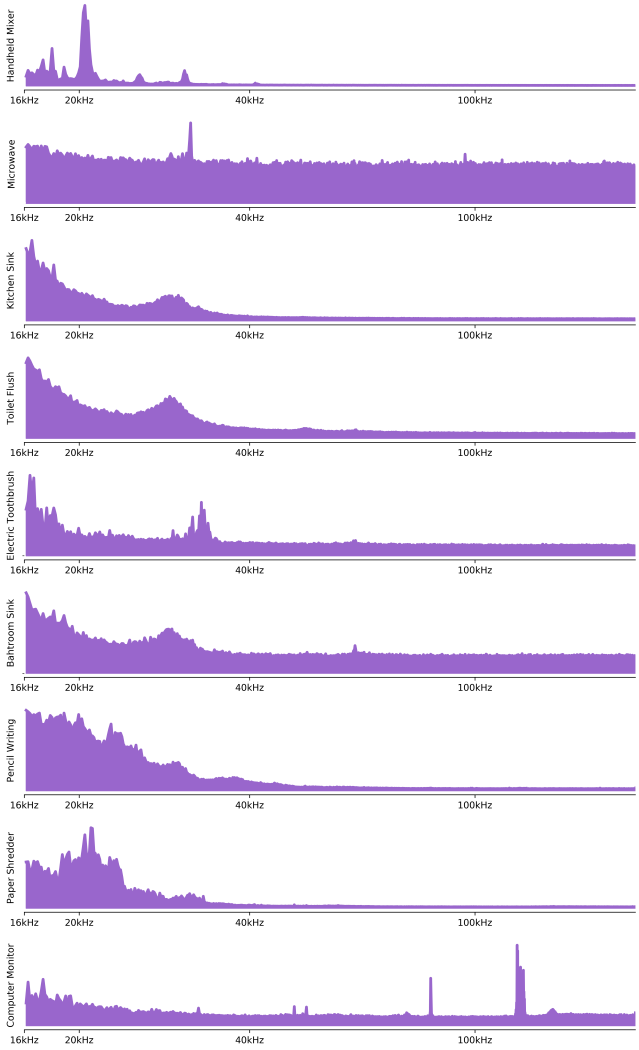

Figure 9: Characteristic FFTs in ultrasound only for each of our selected real world objects.

20% improvement over privacy-preserving audible (where speech is removed). When privacy-preserving audible is combined with ultrasound, classification accuracies surpass traditional audible performances that includes speech frequencies. These two frequency combinations are precisely what PrivacyMic leverages as input when using its speech and audible filters.

There is a trade-off between accuracy and privacy with many systems: reducing the amount of information available to ML models makes it more challenging to perform classification tasks accurately. Fortunately, there is a plentiful amount of information in ultrasound, and re-running our Inaudible Sounds Evaluation with a cutoff of 20kHz (the absolute upper range of hearing in all humans) resulted in only a 5% decrease in accuracy. Many of the most valuable ultrasound bands are significantly outside of human hearing: 5 of the Top 20 bands are >27kHz, and all ultrasonic bands at or below 56kHz are above-average importance. It's also important to note that all ultrasonic bands have non-zero feature importance, meaning they contribute to the classification performance. Thus, while increasing the filter cutoff to improve privacy will affect accuracy, we still expect good performance even when the cutoff is significantly outside of human hearing.

As the number of listening devices grows in our lives, the implications of privacy become of greater importance. All smart speech-based personal assistants require a key-phrase for invocation, like "Hey Siri" or "Ok Google." In an ideal world, these devices do not "listen" until the phrase is said, but, this prohibits a platform from truly achieving real-time, always-running activity recognition. The converse is always listening devices, which are continuously processing sounds. There are serious privacy concerns around these devices, as improper handling of data can lead to situations where speech and sensitive audio data is recorded and preserved. While our eavesdropping evaluation is by no means an exhaustive study to prove that PrivacyMic definitively removes all traces of speech, it shows that at least in the case of someone "listening in" to audio data recorded via PrivacyMic that speech is no longer intelligible.

While current speech recognition systems were also unable to detect speech content in our evaluation, we expect future eavesdropping attacks to more likely be algorithmic rather than a human listening and may be able to perform speech recovery using the remaining speech fragments that humans cannot find intelligible. One possible way to preemptively thwart these attacks is to increase the filter cutoff such that no speech fragments remain. Future work should explore how to definitively safeguard against these algorithmic attacks by training models to eavesdrop on "safe" sounds and validating the safeness of the cutoff. While we cannot prevent every future attack, we argue PrivacyMic presents a minimum baseline for privacy.

Using ultrasonic frequencies also has implications on device hardware. In Figure 4, looking at the ultrasound bins, there's a drop-off in importance for frequency components above 56kHz. Further, all of the ultrasonic bins that appear in the top 20 feature importances (Figure 4) exist outside of the range of most microphones (above 20kHz), yet below 45kHz. While components outside of those ranges are not unimportant, it suggests that future devices are not far away from capturing a few more high-importance frequency ranges before the cost outweighs the benefit. Simply put, if the upper limit of devices were extended from 20kHz to 56kHz,

Unified Model: Speech Filter On

Speech Filtered	Nothing	Mixer	Microwave	K. Sink	Toilet	Toothbrush	B. Sink	Writing	Shredder	Monitor
Nothing	**100.0%**	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Mixer	0.0%	**100.0%**	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Microwave	0.0%	0.0%	**100.0%**	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Kitchen Sink	0.0%	0.0%	0.0%	**99.0%**	1.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Toilet	0.0%	0.0%	0.0%	6.0%	**94.0%**	0.0%	0.0%	0.0%	0.0%	0.0%
Toothbrush	0.0%	0.0%	0.0%	0.0%	1.0%	**99.0%**	0.0%	0.0%	0.0%	0.0%
Bathroom Sink	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	**100.0%**	0.0%	0.0%	0.0%
Writing	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.0%	**99.0%**	0.0%	0.0%
Shredder	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	**100.0%**	0.0%
Monitor	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.0%	1.0%	**98.0%**

Unified Model: Audible Filter On

Audible Filtered	Nothing	Mixer	Microwave	K. Sink	Toilet	Toothbrush	B. Sink	Writing	Shredder	Monitor
Nothing	**95.0%**	0.0%	5.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Mixer	0.0%	**100.0%**	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Microwave	14.0%	0.0%	**84.0%**	0.0%	0.0%	0.0%	2.0%	0.0%	0.0%	0.0%
Kitchen Sink	0.0%	0.0%	0.0%	**100.0%**	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Toilet	0.0%	0.0%	0.0%	11.0%	**89.0%**	0.0%	0.0%	0.0%	0.0%	0.0%
Toothbrush	0.0%	0.0%	0.0%	0.0%	3.0%	**97.0%**	0.0%	0.0%	0.0%	0.0%
Bathroom Sink	0.0%	0.0%	2.0%	0.0%	0.0%	1.0%	**97.0%**	0.0%	0.0%	0.0%
Writing	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.0%	**98.0%**	0.0%	0.0%
Shredder	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	**100.0%**	0.0%
Monitor	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.0%	1.0%	**98.0%**

Figure 10: Confusion matrices for PrivacyMic's real world evaluation for the two unified models, with speech filtered out (left) and audible filtered out (right).

they would capture 86.4% of the total feature importance of the full spectrum analyzed in this study.

Further, using inaudible frequencies encompass sensing capabilities that were commonly associated with other sensors. For example, to determine whether the lights or a computer monitor is on, a photosensor and RF module are reasonable choices of sensors. Utilizing ultrasound, PrivacyMic now can "hear" light bulbs and monitors, two devices that are silent to humans. In the case of the computer monitor, real-time classification can be seen in the Video Figure.

9 CONCLUSION

In conclusion, our work explored inaudible sources as feature sources for acoustic activity recognition. We compiled a first-of-its-kind dataset, containing 127 items and 6.2 hours of audio data collected across seven buildings. With this data, we performed a spectral information power analysis, showing that ultrasound frequencies comprise five of the top ten features and has the highest average feature importance compared to audible and infrasound frequencies. We then simulated the classification performance of our device using this collected dataset and found privacy-preserving accuracies of up to 91.4%, compared to 50.5% when using speech-filtered audible alone. We used these findings to design PrivacyMic's in-hardware speech and audible filters. We used these filters to perform an eavesdropping evaluation, where participants were asked to listen to audio clips of filtered speech: none of the participants were able to transcribe a single word. We then performed a real-world evaluation across three common environments, kitchen, bathroom, and office, and found greater than 95% accuracy across all three environments and filter settings. Inaudible sounds are clearly a rich source of signals and we hope our findings inspire more work and wider adoption of these privacy-preserving frequency bands for acoustic activity recognition.

REFERENCES

[1] Amazon.com. 2019. Pioneer SP-BS22-LR. Website. Retrieved September 20, 2019 from https://www.amazon.com/Pioneer-SP-BS22-LR-Designed-Bookshelf-Loudspeakers/dp/B008NCD2LG.

[2] D. Arp, E. Quiring, C. Wressnegger, and K. Rieck. 2017. Privacy Threats through Ultrasonic Side Channels on Mobile Devices. In *2017 IEEE European Symposium on Security and Privacy (EuroS P)*. 35–47.

[3] Avisoft. 2019. Knowles FG Response Curve. Website. Retrieved September 20, 2019 from https://www.avisoft.com/usg/KnowlesFGO.htm.

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*. 892–900.

[5] BlueDesigns. 2019. Blue Yeti Microphone. Website. Retrieved September 20, 2019 from https://www.bluedesigns.com/products/yeti/#.

[6] Emre Cakır, Giambattista Parascandolo, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. 2017. Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 6 (2017), 1291–1303.

[7] Ray-I Chang, Chien-Chang Huang, Liang-Bin Lai, and Chia-Yun Lee. 2018. Query-Based Machine Learning Model for Data Analysis of Infrasonic Signals in Wireless Sensor Networks. In *Proceedings of the 2Nd International Conference on Digital Signal Processing* (Tokyo, Japan) *(ICDSP 2018)*. ACM, New York, NY, USA, 114–118. https://doi.org/10.1145/3193025.3193031

[8] Dodotronic. 2019. Ultramic384K. Website. Retrieved September 20, 2019 from https://www.dodotronic.com/ultramic384k/.

[9] Benjamin Elizalde, Rohan Badlani, Ankit Shah, Anurag Kumar, and Bhiksha Raj. 2018. Nels-never-ending learner of sounds. (2018). arXiv:1801.05544

[10] Antti J Eronen, Vesa T Peltonen, Juha T Tuomi, Anssi P Klapuri, Seppo Fagerlund, Timo Sorsa, Gaëtan Lorho, and Jyri Huopaniemi. 2005. Audio-based context recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 1 (2005), 321–329.

[11] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. 2015. Reliable detection of audio events in highly noisy environments. *Pattern Recognition Letters* 65 (2015), 22–28.

[12] Frederic Font, Gerard Roma, and Xavier Serra. 2013. Freesound Technical Demo. In *Proceedings of the 21st ACM International Conference on Multimedia* (Barcelona, Spain) *(MM '13)*. ACM, New York, NY, USA, 411–412. https://doi.org/10.1145/2502081.2502245

[13] Jon E. Froehlich, Eric Larson, Tim Campbell, Conor Haggerty, James Fogarty, and Shwetak N. Patel. 2009. HydroSense: Infrastructure-mediated Single-point Sensing of Whole-home Water Activity. In *Proceedings of the 11th International Conference on Ubiquitous Computing* (Orlando, Florida, USA) *(UbiComp '09)*. ACM, New York, NY, USA, 235–244. https://doi.org/10.1145/1620545.1620581

[14] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 776–780.

[15] Pascal Getreuer, Chet Gnegy, Richard F Lyon, and Rif A Saurous. 2017. Ultrasonic communication using consumer hardware. *IEEE Transactions on Multimedia* 20, 6 (2017), 1277–1290.

[16] Mayank Goel, Elliot Saba, Maia Stiber, Eric Whitmire, Josh Fromm, Eric C. Larson, Gaetano Borriello, and Shwetak N. Patel. 2016. SpiroCall: Measuring Lung Function over a Phone Call. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 5675–5685. https://doi.org/10.1145/2858036.2858401

[17] Sidhant Gupta, Matthew S. Reynolds, and Shwetak N. Patel. 2010. ElectriSense: Single-point Sensing Using EMI for Electrical Event Detection and Classification in the Home. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing* (Copenhagen, Denmark) *(UbiComp '10)*. ACM, New York, NY, USA, 139–148. https://doi.org/10.1145/1864349.1864375

[18] Infiltec. 2019. Infiltec INFRA20 Infrasound Sensor. Website. Retrieved September 20, 2019 from http://www.infiltec.com/Infrasound@home/.

[19] Yasha Iravantchi, Mayank Goel, and Chris Harrison. 2019. BeamBand: Hand Gesture Sensing with Ultrasonic Beamforming. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI*

'19). ACM, New York, NY, USA, Article 15, 10 pages. https://doi.org/10.1145/3290605.3300245

[20] Yasha Iravantchi, Yang Zhang, Evi Bernitsas, Mayank Goel, and Chris Harrison. 2019. Interferi: Gesture Sensing Using On-Body Acoustic Interferometry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 276, 13 pages. https://doi.org/10.1145/3290605.3300506

[21] KeySight. 2019. Agilent 33521A Function Generator. Website. Retrieved September 20, 2019 from https://www.keysight.com/en/pd-1871159-pn-33521A/function-arbitrary-waveform-generator-30-mhz?&cc=US&lc=eng.

[22] Mone Kijima, Yuta Miyagawa, Hayato Oshita, Norihisa Segawa, Masato Yazawa, and Masa-yuki Yamamoto. 2018. Multiple door opening/closing detection system using infrasound sensor. In *Proceedings of the 17th ACM/IEEE International Conference on Information Processing in Sensor Networks*. IEEE Press, 126–127.

[23] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-Play Acoustic Activity Recognition. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. ACM, New York, NY, USA, 213–224. https://doi.org/10.1145/3242587.3242609

[24] Gierad Laput, Yang Zhang, and Chris Harrison. 2017. Synthetic Sensors: Towards General-Purpose Sensing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 3986–3999. https://doi.org/10.1145/3025453.3025773

[25] Eric C. Larson, TienJui Lee, Sean Liu, Margaret Rosenfeld, and Shwetak N. Patel. 2011. Accurate and Privacy Preserving Cough Sensing Using a Low-Cost Microphone. In *Proceedings of the 13th International Conference on Ubiquitous Computing* (Beijing, China) *(UbiComp '11)*. Association for Computing Machinery, New York, NY, USA, 375–384. https://doi.org/10.1145/2030112.2030163

[26] S. Lee, E. Nemati, and J. Kuang. 2018. Configurable Pulmonary-Tuned Privacy Preservation Algorithm for Mobile Devices. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 1107–1112.

[27] Lydia Manikonda, Aditya Deotale, and Subbarao Kambhampati. 2018. What's up with Privacy? User Preferences and Privacy Concerns in Intelligent Personal Assistants. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*. Association for Computing Machinery, New York, NY, USA, 229–235. https://doi.org/10.1145/3278721.3278773

[28] Jess McIntosh, Asier Marzo, Mike Fraser, and Carol Phillips. 2017. EchoFlex: Hand Gesture Recognition Using Ultrasound Imaging. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 1923–1934. https://doi.org/10.1145/3025453.3025807

[29] Bjoern H Menze, B Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A Hamprecht. 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics* 10, 1 (2009), 213.

[30] The Cornell Lab of Ornithology. 2019. Elephant Listening Project: Deep into Infrasound. Website. Retrieved September 20, 2019 from https://elephantlisteningproject.org/all-about-infrasound/.

[31] Future of Privacy Forum. 2020. "Always On: Privacy Implications of Microphone-Enabled Devices". Website. Retrieved Sept 17, 2020 from https://www.ftc.gov/system/files/documents/public$_c$omments/2016/08/00003-128652.pdf.

[32] PUIAudio. 2019. PUI Audio AB1290B. Website. Retrieved September 20, 2019 from http://www.puiaudio.com/product-detail.aspx?categoryId=5&partnumber=AB1290B.

[33] Dale Purves. 2001. *The Audible Spectrum*. Sunderland Associates, Sunderland, MA, USA.

[34] Rolf Quam, Ignacio Martínez, Manuel Rosa, Alejandro Bonmatí, Carlos Lorenzo, Darryl J. de Ruiter, Jacopo Moggi-Cecchi, Mercedes Conde Valverde, Pilar Jarabo, Colin G. Menter, J. Francis Thackeray, and Juan Luis Arsuaga. 2015. Early hominin auditory capacities. *Science Advances* 1, 8 (2015). https://doi.org/10.1126/sciadv.1500355 arXiv:https://advances.sciencemag.org/content/1/8/e1500355.full.pdf

[35] RecordingHacks. 2019. Blue Yeti Microphone Response Curve. Website. Retrieved September 20, 2019 from http://recordinghacks.com/microphones/Blue-Microphones/Yeti.

[36] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. 2017. BackDoor: Making Microphones Hear Inaudible Sounds. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services* (Niagara Falls, New York, USA) *(MobiSys '17)*. Association for Computing Machinery, New York, NY, USA, 2–14. https://doi.org/10.1145/3081333.3081366

[37] Yvan Saeys, Thomas Abeel, and Yves Van de Peer. 2008. Robust feature selection using ensemble feature selection techniques. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 313–325.

[38] Florian Sainjeon, Sébastien Gaboury, and Bruno Bouchard. 2016. Real-Time Indoor Localization in Smart Homes Using Ultrasound Technology. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (Corfu, Island, Greece) *(PETRA '16)*. ACM, New York, NY, USA, Article 6, 4 pages. https://doi.org/10.1145/2910674.2910718

[39] Valkyrie Savage, Andrew Head, Björn Hartmann, Dan B. Goldman, Gautham Mysore, and Wilmot Li. 2015. Lamello: Passive Acoustic Sensing for Tangible Input Components. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. ACM, New York, NY, USA, 1277–1280. https://doi.org/10.1145/2702123.2702207

[40] ShotSpotter. 2019. ShotSpotter Gunshot Detection, Location and Forensic Analysis. Website. Retrieved September 20, 2019 from https://www.shotspotter.com.

[41] Takeshi Sugawara, Benjamin Cyr, Sara Rampazzi, Daniel Genkin, and Kevin Fu. 2020. Light Commands: Laser-Based Audio Injection Attacks on Voice-Controllable Systems. In *29th USENIX Security Symposium (USENIX Security 20)*.

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

[43] Wikipedia. 2020. "Wikipedia: Hearing range". Website. Retrieved Sept 17, 2020 from https://en.wikipedia.org/wiki/Hearing$_r$ange.

[44] Wikipedia. 2020. "Wikipedia: Mel scale". Website. Retrieved Sept 17, 2020 from https://en.wikipedia.org/wiki/Mel$_s$cale.

[45] Cheng Zhang, Qiuyue Xue, Anandghan Waghmare, Ruichen Meng, Sumeet Jain, Yizeng Han, Xinyu Li, Kenneth Cunefare, Thomas Ploetz, Thad Starner, Omer Inan, and Gregory D. Abowd. 2018. FingerPing: Recognizing Fine-grained Hand Poses Using Active Acoustic On-body Sensing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 437, 10 pages. https://doi.org/10.1145/3173574.3174011

[46] Hanbin Zhang, Chen Song, Aosen Wang, Chenhan Xu, Dongmei Li, and Wenyao Xu. 2019. PDVocal: Towards Privacy-Preserving Parkinson's Disease Detection Using Non-Speech Body Sounds. In *The 25th Annual International Conference on Mobile Computing and Networking* (Los Cabos, Mexico) *(MobiCom '19)*. Association for Computing Machinery, New York, NY, USA, Article 16, 16 pages. https://doi.org/10.1145/3300061.3300125