# Evaluating the Effectiveness of Protection Jamming Devices in Mitigating Smart Speaker Eavesdropping Attacks Using Gaussian White Noise

Payton Walker
Texas A&M University
College Station, Texas, USA
prw0007@tamu.edu

Nitesh Saxena
Texas A&M University
College Station, Texas, USA
nsaxena@tamu.edu

## ABSTRACT

Protection Jamming Devices (PJD) are specialized tools designed to sit on top of virtual assistant (VA) smart speakers and hinder them from "hearing" nearby user speech. PJDs aim to protect you from eavesdropping attacks by injecting a jamming signal directly into the microphones of the smart speaker. However, current signal processing routines can be used to reduce noise and enhance speech contained in noisy audio samples. Therefore, we identify a potential vulnerability for speech eavesdropping via smart speaker recordings, even when a PJD is being used. If an attacker can gain access to or facilitate smart speaker recordings they may be able to compromise a user's speech with successful noise cancellation. Specifically, we are interested in the potential for Gaussian white noise (GWN) to be an effective jamming signal for a PJD. To our knowledge, the effectiveness of white noise and PJDs to protect against eavesdropping attacks has yet to receive a systematic evaluation that includes physical experiments with an actual PJD implementation.

In this work we construct our own PJD, specialized for consistent experimentation, to simulate an attack scenario where recordings from a smart speaker, in the presence of normal speech and the PJDs jamming signal, are recovered. We perform substantial data collection under different settings to build a repository of 1500 recovered audio samples. We applied post-processing on our dataset and conducted an extensive signal/speech quality analysis including both time and frequency domain inspection, and evaluation of metrics including cross-correlation, SNR, and PESQ. Lastly, we performed feature extraction (MFCC) and built machine learning classifiers for tasks including speech (digit) recognition, speaker identification, and gender recognition. We also attempted song recognition using the Shazam app. For all speech recognition tasks that we attempted, we were able to achieve classification accuracies above that of random guessing (46% for digit recognition, 51% for speaker identification, 80% for gender identification), as well as demonstrate successful song recognition. These results highlight the real potential for attackers to compromise user speech, to some extent, using smart speaker recordings; even if the smart speaker is protected by a PJD.


*ACSAC '21, December 6–10, 2021, Virtual Event, USA*

ACM ISBN 978-1-4503-8579-4/21/12…$15.00
https://doi.org/10.1145/3485832.3485896

## CCS CONCEPTS

• **Security and privacy → Privacy protections**; *Privacy-preserving protocols*; *Hardware-based security protocols*;

## KEYWORDS

speech masking, jamming, eavesdropping, white noise

**ACM Reference Format:**
Payton Walker and Nitesh Saxena. 2021. Evaluating the Effectiveness of Protection Jamming Devices in Mitigating Smart Speaker Eavesdropping Attacks Using Gaussian White Noise. In *Annual Computer Security Applications Conference (ACSAC '21), December 6–10, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3485832.3485896

## 1 INTRODUCTION

Voice Controllable Systems (VCS), and specifically smart speaker devices, have gained significant popularity in home and business environments throughout recent years. Today, at least 35% of the U.S. population (18+) owns a smart speaker; and that number is expected to increase to 75% by 2025 [30]. The study, conducted by NPR and Edison Research, also found the average smart speaker household had multiple (2.6) devices [7]. These smart speakers can be fully interfaced via vocal commands which introduces a new form of accessibility. This allows certain user groups to utilize functions that otherwise may not be possible for them (i.e., due to physical disability). Major companies like Amazon and Google have released different versions of their own standalone VCS smart speakers, many being relatively inexpensive. Models such as the Amazon Echo Dot [3] and Google Home Mini [4] are becoming a common commodity because of their low cost and their ability to connect other smart devices (i.e., thermostats, locks, etc.).

Due to their amassed popularity, the security and privacy of user data, particularly their speech, has become a major concern. Many people believe that these smart speaker devices can be used by malicious attackers, the government, or even the companies selling the speakers to eavesdrop on their users at any point. These concerns have led to significant news and media coverage describing the potential for such attacks [5, 22, 23, 28, 31]. The implications of smart speaker eavesdropping could be devastating if we consider the sensitive environments they could be placed in (i.e., home, office, etc.). In these settings there may be a lot of confidential speech from the user that should remain private. And at face value, the potential for smart speaker eavesdropping may seem high because of the "always on" nature of the microphones for detecting the wake word. Although Google and Amazon report that their devices do

not record user conversations [2, 6], many people are not convinced and believe the threat is real.

In response, recent studies and projects have begun to explore defensive techniques (namely microphone jamming) to mitigate these eavesdropping attacks. These works implement microphone jamming in what is called Protection Jamming Devices (PJD). PJDs use tiny speakers housed in a mount, and are designed to rest on top of the user's smart speaker device. The jamming device will play some type of noise (i.e., white noise, chatter, ultrasonic) through the tiny speakers, directly into the microphones of the smart speaker, in order to block any speech in the surrounding area. These devices are always on and continue to jam the audio input of a smart speaker until a wake word unique to the PJD is detected (using its own inbuilt microphone). When the new wake word is detected the PJD activates the smart speaker and stops jamming its microphones so the user's command can be processed.

In recent years we have seen new attention given to this type of solution. New PJDs like Project Alias [14] and Paranoid's Home Wave [8] are currently on the market and available to the public. Additionally, MicShield [41] is an academic paper that presents a PJD solution using ultrasonic noise for jamming. These products and research demonstrate a new defense against smart speaker eavesdropping. If this approach can be verified as completely effective at stopping eavesdropping attacks, it could introduce a new sense of security and protection for current and new smart speaker users. Further, these devices are not expensive meaning they could easily be adopted by existing users, or the jamming technology could potentially be integrated in future models of the popular smart speaker brands. However, this all depends on the PJD's ability to produce a jamming signal that can 1) interfere with the microphone's ability to detect nearby speech, and 2) does not bother or annoy the user (e.g., transparent to them). Jamming using ultrasonic noise has the benefit of being inherently undetectable by the human ear, and it can mechanically hinder a microphone from recording. But in the case of audible noise as the jamming signal (white or chatter), finding an adequate loudness for the noise can be a delicate task. Chatter noise in particular faces additional challenges because the noise contains a more dynamic and recognizable combination of sounds that would be more distracting than a static white noise. Additionally, ultrasonic jamming does not stop the microphone from recording, but rather obfuscates the speech that is recorded beyond the point of recognition. Therefore, the potential for speech recovery still exists because of signal processing techniques that could potentially remove the injected jamming noise and reveal the original speech. While chatter type noise has been successfully implemented in the Project Alias PJD solution, we chose to investigate white noise in this initial study because of its popularity in other current speech masking solutions [1, 9, 11]. Further, to our knowledge the effectiveness of white noise injected in the foreground of audio recordings, for masking speech, has yet to be explored with physical experiments ([41] only simulated white noise jamming).

In this work we study the efficacy of Protection Jamming Devices that use audible Gaussian White noise for the jamming signal for mitigating smart speaker eavesdropping attacks. We build our own PJD implementation (designed for experimentation) and conduct experiments that expose a smart speaker (Amazon Echo Dot) to speech audio and the jamming noise. The recordings were saved from the Alexa Voice History and processed using off-the-shelf noise cancellation and speech enhancement routines. We extracted different features from our samples and built classifiers to attempt speech, speaker, and gender recognition. This attack model is designed to simulate a less-sophisticated, real-world attacker in order to observe a baseline for attack success. Also, using off-the-shelf techniques makes the attack model more *practical* and *accessible* to even low-capability attackers. Our results suggest that speech contained in smart speaker recordings, during active GWN jamming, can be compromised. Further, we believe attack success can increase with more sophisticated and skilled attackers.

**Contributions:** The main contributions made in this work are summarized below:

(1) We provide an overview of existing PJD implementations and other related works (Section 2).
(2) We build our own PJD device modeled after existing implementations and conduct experiments to build a dataset of smart speaker recordings of speech in the presence of a jamming signal (Section 4).
(3) We performed an extensive signal/speech quality analysis including time and frequency domain inspection, and using quality metrics such as cross-correlation, SNR, and PESQ (Section 6).
(4) Lastly, we used machine learning to attempt speech (digit) recognition, speaker and gender identification; as well as attempt song recognition. We achieve classification accuracies better than random guessing (46% for digit recognition, 51% for speaker identification, 80% for gender identification); and demonstrate successful song recognition. Our results highlight a potential point of vulnerability in PJDs that use acoustic jamming signals (Section 7).

The significance of this study is that it systematically confirms, in an academic setting, that standard jamming noises such as white noise are not effective at protecting user speech from even unsophisticated attackers that only use standard off-the-shelf signal processing techniques. Existing PJDs such as Project Alias [14] and Home Wave [8] do not use the standard white noise giving them more success at masking user speech, and at the least they significantly decrease the level at which speech can be compromised by an attacker (increasing the difficulty of the attack). However, we show that compromising user speech may be successful to some extent with effective noise cancellation and speech enhancement routines for processing the noisy audio. Therefore, as signal processing techniques continue to improve, PJD devices must continue to accommodate for an attacker's increased ability in order to remain an effective defensive strategy. This work makes no claims about the effectiveness of the existing PJD solutions that do not use Gaussian white noise [8, 14, 41]. They are only used to inform the design of our own jamming device (hardware and software).

## 2 BACKGROUND

**Protection Jamming Device:** Recent projects by researchers and independent developers have produced a new mechanism to protect against smart speaker eavesdropping, called Protection Jamming Devices (PJD). They use tiny speakers housed in a mount, and are designed to rest on top of the user's smart speaker device.

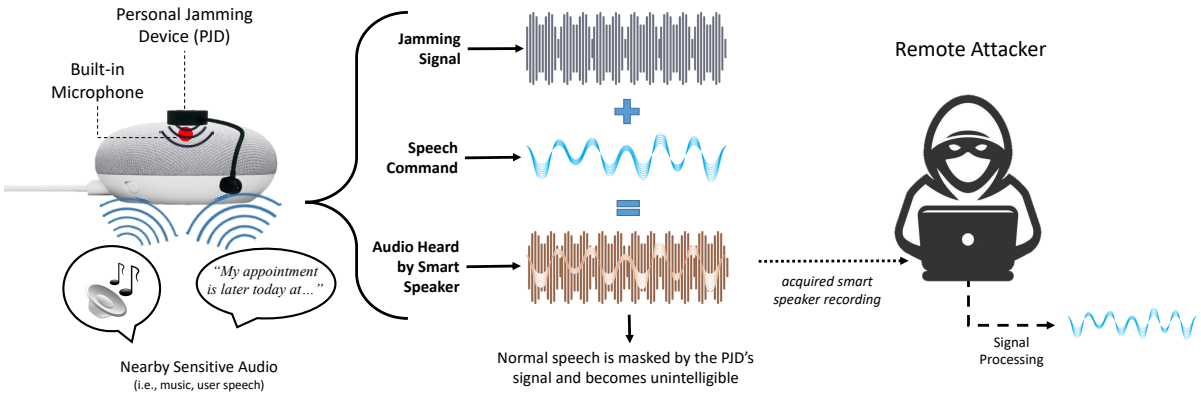



**Figure 1: Depiction of how a PJD functions to mask sensitive user speech; and how a remote attacker could potentially eavesdrop by removing the injected noise to recover the original speech.**

The PJD will play some type of jamming signal, directly into the smart speaker microphones, in order to block any speech in the surrounding area. Figure 1 illustrates the basic function of a PJD, as well as the potential threat faced if an attacker can compromise the smart speaker's recordings. These devices are always on and continue to jam the audio input of a smart speaker until a wake word unique to the PJD is detected. When it hears the new wake word, the PJD activates the smart speaker and stops jamming so the user's command can be heard and processed.

In the market, we have seen a few PJD products become available in recent years. Project Alias [14] is a device that was created in an independent project by Bjorn Karmann, and made open-source to the public. Home Wave [8] was created by the company Paranoid Inc. and is available for purchase on their website. Both of these devices have similar setups using tiny speakers attached to a housing that rests on top of a smart speaker. The speakers play an acoustic jamming signal (at a low loudness that remains undetectable to the user). Additionally, both devices are equipped with their own microphone for detecting a unique wake word. Project Alias can be trained to respond to any word, and the Home Wave device recognizes "Paranoid" as the wake word. These new devices have already been featured many times in news and media (Project Alias - [24, 36, 38], Home Wave - [12, 16, 20]) demonstrating the popularity of devices that can offer increased privacy.

**Acoustic vs. Ultrasonic Noise:** Although jamming with ultrasonic noise can be very effective at blocking any snooping devices from eavesdropping in a specific area, solutions using acoustic jamming noises ([8, 14]) are more cost effective for the average user. Additionally, research has shown that prolonged exposure to ultrasonic noise can have harmful effects on humans including noise induced hearing loss [39] and loss of concentration [26]. Therefore, smart speaker jamming devices that utilize ultrasonic noise may not be viable as a long-term, in-home solution for many users. Because of this, we feel it is still important to understand the limitations of jamming solutions that use *audible* noise.

Additionally, in a more generalized study by Cheng et al. [19], the authors evaluate the effectiveness of different jamming signals and the effect of wake word and noise overlap on jamming success. Their work demonstrated that Gaussian white noise can be used for successful jamming under certain conditions (e.g., with a strong audio signal (10dB SNR)). However, their evaluation used programmatic signal injection to simulate a speech masking scenario. Additionally, the normal speech and jamming signal were combined into one signal before being fed to ASR in their experiments. Our work differs because it looks to assess the masking potential of GWN in a physical PJD implementation. This allows us to observe the real-world limitations, if any, of injecting noise into a smart speaker device for blocking nearby speech.

**Noise Cancellation:** In order to increase speech recognition potential, current smart speakers will instantly process audio input to try and enhance any contained speech. Specifically, intricate noise cancellation may be applied to remove any unnecessary sounds from the audio file, before running it through automatic speech recognition. For example, when the Echo Dot 2 (used in our study) receives audio input, it immediately transmits it to the Alexa Voice Service (AVS) on the internet. Available research from the Amazon Group reveal that processes such as adaptive linear filtering and acoustical echo cancellation [47], adaptive beamforming [46], and spacial localization [42] utilizing the multi-microphone array are occurring during this time. Uniquely, the noise injection technique of the PJD can introduce challenges for the existing noise cancellation routines, reflecting their potential for success. The existing signal processing techniques are not equipped to handle audio input where the noise source is in the foreground, and the normal speech is farther away. This is what allows PJD devices to be successful at hindering the ASR function of a smart speaker. However, if we consider the potential a human attacker could have with ample time and access, and improved signal processing techniques, it is unlikely a foreground injected jamming signal can remain effective.

**Related Works:** A device that similarly uses speakers at a close distance to the smart speaker microphones, for jamming signal injection, is MicShield [41] that was presented in an academic work by Sub et al. This device differs from the first two in that it uses

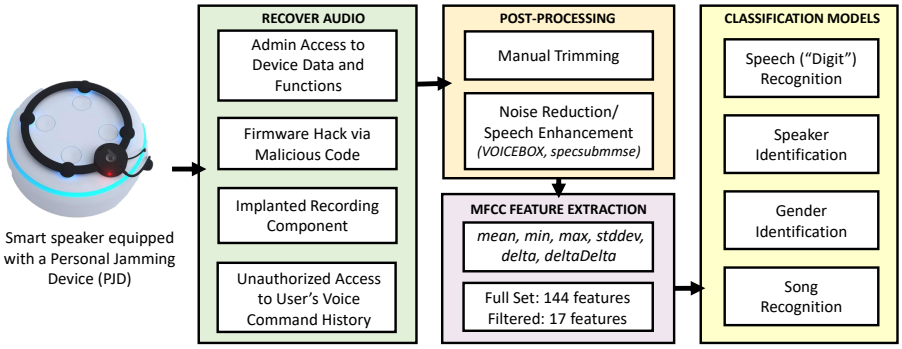



**Figure 2: Diagram depicting an attack model that can be used to target our defined threat model.**

ultrasonic sounds for the jamming signal which affects the operations of the microphone, as opposed to simply masking the recorded audio in noise. The authors performed some preliminary testing using white noise and found that a very low SNR (-15dB) can be effective for the PJD setup. They determined speech at 75 dB could be jammed with a white noise signal at 90 dB; which is not a viable loudness for a PJD using an audible jamming noise (because of user disturbance). Our work further develops this insight to assess the effectiveness of white noise jamming at loudness levels acceptable to a user, through a formal academic study.

The use of ultrasonic noise is also seen in other defenses such as the Patronus system [34]. Patronus generates low-frequency noise called a "scramble" that exploits the nonlinear effects of smart speaker microphones to prevent unauthorized recordings and improve the quality of authorized recordings. Another work by Chen et al. [18] presents a wearable bracelet composed of ultrasonic speakers that can disable microphones that are near the user wearing it. Rather than targeting a specific VCS device, this bracelet is intended to disrupt recording from all microphones close to the user (i.e., smartphones, smartwatches, etc.). We have even seen an artistic, non-technical defense approach similar to a physical barrier. May Safwat designed and constructed a bust of the whistleblower Edward Snowden that contains a hollow copper tube. The bust is simply meant to sit over a smart speaker (completely covering the top and all sides) in order to block any a audio or wireless signal communication (i.e., internet) with the smart speaker. Clearly, this holds the same level of protection as simply powering off/unplugging the smart speaker device (e.g., cause complete DoS). But it still speaks to the growing interest surrounding these type of devices.

Aside from smart speaker eavesdropping defenses, speech masking techniques for protecting live human speech have been explored that still reveal new knowledge about the potential of speech masking/jamming. In a work by Phunruangsakao et al. [37], the authors develop a scheme to ensure speech privacy by limiting the Speech Transmission Index (STI). Instead of measuring the Room Impulse Response (RIR), the authors estimate the STI and feed it to an RIR model. Kim et al. [32] explores the potential of Active Noise Cancellation (ANC) for increasing speech privacy. Their work demonstrates that ANC in a specific direction can decrease the need for a masking signal, and the same masking affect can be achieved with a lower masking signal volume (5 dB lower). The authors evaluate speech intelligibility in their analysis using Speech Intelligibility Index (SII) and Speech Reception Threshold (SRT). Lastly, research conducted by Krasnov et al. [33] looks to mask an original speech signal (with no user disturbance) by targeting key components of the speech that carry information required for recognition. Both amplitude and temporal smearing techniques are used to generate a modified masking noise to negate some effects of reverberation and increase speech privacy. Our work uniquely builds on these existing works by performing the first evaluation of the presented mechanisms, using a standard white noise jamming signal, against a simple, but realistic eavesdropping attack.

## 3 THREAT MODEL

For our threat model we consider a user environment that contains a VCS smart speaker device equipped with a PJD. This could be a personal device placed in a user's home, or a work tool used in their place of business. In these scenarios, the smart speakers are exposed to sensitive speech that the users need to keep confidential. The attacker in our threat model seeks to compromise or eavesdrop a victim user's speech and are able to acquire (noisy) recordings from the victim's smart speaker device. For example, an employee of the device's shipping company could implant a component or inject malicious code in the firmware that would allow them to acquire user speech recordings. They can perform standard post-processing techniques on the victim's speech recordings including noise cancellation and speech enhancement. Lastly, the attacker possesses the machine learning knowledge needed to build a speech recognition (or other related speech task) model. Specifically, the attacker in our threat model looks to achieve *speech (digit) recognition*, *speaker identification*, *gender identification*, and *song recognition*. Each of these speech tasks can reveal sensitive information about a user that an attacker can abuse or even sell to companies for things like targeted advertising. Speech recognition can reveal actual speech content, while the other tasks like song recognition may reveal a user's personal and private interests. Figure 2 depicts a model an attacker may use, and we recreate experimentally in this study, to eavesdrop on speech via compromised smart speaker recordings.

We consider digit recognition because it represents the potential for PIN/password/account# leakage which would be very devastating if acquired by a malicious attacker. If we consider the scenario

of making a purchase over the phone, *which is becoming increasingly more common during this pandemic as people are encouraged to quarantine at home*, we can see an instance where a user may vocalize sensitive numerical information (i.e., credit card number and security code) near their smart speaker. Additionally, speaker identification could reveal to an attacker how many people live in a home, as well as when particular people are using the smart speaker and what they use it for. This information could be sold to companies that will use it for targeted advertising, or for something even more malicious such as a home robbery. Similarly, Gender identification may also be used for targeted advertisements, or to help the eavesdropper understand who lives in the home.

In a real-world scenario, there are a few potential attackers that can attempt eavesdropping in this way. Mainly, the companies providing these smart speaker devices, and their employees, are in a unique position to perform such an attack. The major companies like Amazon and Google could easily program these devices to record and transmit user audio to company servers, or install a function to allow full access and control of the microphones equipped on the device by an administrator at the company. Although the companies claim this does not occur, many people are still highly suspicious of this possibility [5, 28]. Aside from the companies that sell the devices, this threat model is also applicable to any attacker that can gain access to, or even force, recordings from a user's smart speaker. This could be a malicious person with IoT device hacking skills, or even a military entity. Fear of the military exploiting these VA devices is another concern of many people [13, 40, 43].

Similarly, there is growing potential for law enforcement to obtain smart speaker recordings during their investigations for evidence in legal cases [27]. In New Hampshire, a judge ordered Amazon to turn over two days worth of Alexa recordings from the personal smart speaker device in a murder victim's home [45]. The prosecution looked to find evidence leading to the killer in those recordings. Another example occurred in Hallandale Beach, Florida where smart speaker recordings were subpoenaed to reveal evidence of an argument between a murder victim and the prime suspect [25]. Stories like these could be another motivator for some users to invest in a PJD; wanting their conversations kept private. However, as our work demonstrates, significant information can still be recovered from audio samples masked by a PJD which may interest law enforcement. Even simple data such as number of speakers or the speakers' gender could be useful in an investigation.

# 4 EXPERIMENTS & DATA COLLECTION

## 4.1 PJD Implementation

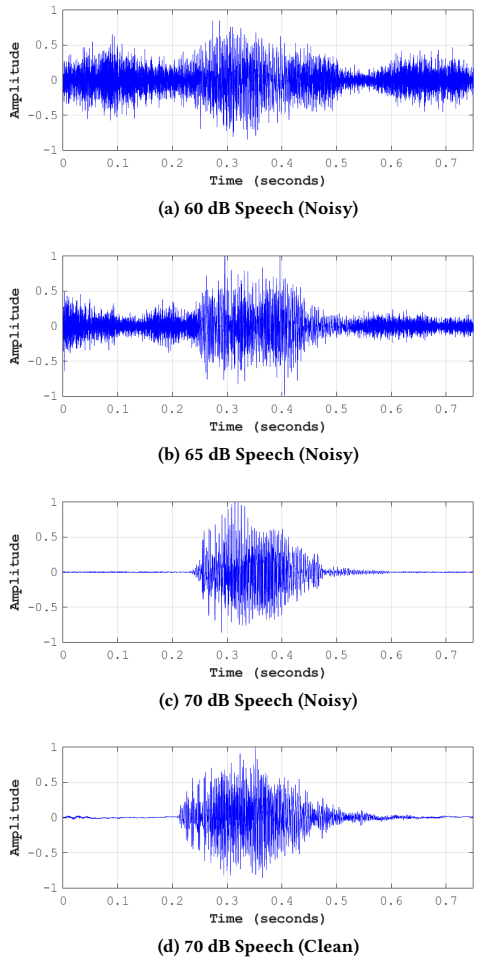
We built our own implementation of a Protection Jamming Device (PJD) based on characteristics of existing PJDs available today. The build instructions and necessary software for the Project Alias device are open source and available online [15]; so we use these materials as the building blocks for of our own device. Like Project Alias, our device uses a Raspberry Pi3 equipped with an SD card and the ReSpeaker 2-Mics Pi HAT expansion board. Additionally, we used a JST 2.0 connector to connect a 16mm tiny speaker. For our jamming signal, we chose to use a standard Gaussian white noise (GWN) that has a flat spectral density and encompasses the 0-8 kHz frequency range. We chose GWN because it is a popular choice for a masking noise and we believe it is a good option for this first academic study in PJD effectiveness.

Unlike the Project Alias device, our implementation does not utilize the 3D-printed shell to house all of the components. Also, we position the tiny speaker directly on top of the center microphone inspired by the design of Home Wave [8]. This will directly inject the jamming noise as audio input into the smart speaker. For the purposes of our experiments, we adapted the Project Alias source the source code to use the GWN jamming signal and added the ability to manually start and stop of the noise using a button on the ReSpeaker expansion board. *These modifications were made so that our PJD could be used for controlled experiments. We are not presenting our constructed device as a new or viable PJD implementation because the design choices we made, while useful for conducting consistent and controlled experiments, add a requirement of user interaction that an actual PJD solution would not have.*

**Determining Injected Noise Volume** Before beginning our experiments, we confirm that our implementation can function successfully as a PJD. We manually adjust the volume of the noise coming from our PJD until it is barely undetectable by a nearby user (confirmed by lab members). In a real-world implementation, the consistent noise played from a PJD cannot be so loud that it disturbs a user in the same space. This is why we adjust the volume level of our PJD jamming signal to a point that is barely detectable by the human ear when they are sitting 0.5 meters from the device. We reason that in a real-world implementation of the device that uses a printed casing to house the speaker and other components, the presence of this noise in the environment will be even lower than what is accepted in our study. With this parameter set, we perform initial testing and confirm the noise injected in the foreground can hinder a smart speaker from recognizing the wake word. This is the key function of PJDs which operate under the assumption that the injected noise will fully mask any nearby user speech.

## 4.2 Experimental Setup

To study how effective our PJD is at masking user speech, we design an experimental setting that exposes an Amazon Echo Dot to both normal speech and the injected noise from our device. Specifically, we attach the tiny speaker of our PJD on top of the center microphone of the Echo Dot, with the Raspberry Pi components sitting next to it. We use an SRS-XB2 Bluetooth speaker to play the speech samples. The Bluetooth speaker is pointed towards and placed approximately 0.5 meters away from the Echo Dot. We test different SPL (dB) levels for the normal speech in our experiments including speech in the normal range for human conversation (60 dB), slightly louder speech (65 dB), as well as very loud speech that is similar to presentation style speaking (70 dB). The SPL for each setting was measured at the smart speaker's location using a digital sound level meter. We test these different SPL levels to generalize our investigation of the effectiveness of GWN, injected in the foreground, to mask nearby speech. In terms of speaker distance, we know that SPL decreases when the distance is doubled by -6 dB. So, if we consider our loudest source speech (70 dB) at 0.5 meters from the Echo Dot; the other SPL levels (65 and 60 dB) would represent distances of about 1 meter and 2 meters, respectively, if we simply moved the 70 dB speech source location.




(a) 60 dB Speech (Noisy)

(b) 65 dB Speech (Noisy)

(c) 70 dB Speech (Noisy)

(d) 70 dB Speech (Clean)

**Figure 3: Time domain graphs generated from post-processed, Alexa recorded samples of speaker FAC saying the digit "One", for each speech SPL tested (60, 65, 70 dB); as well as the time domain graph of a clean sample of the same speech from FAC.**

## 4.3 Data Collection

We performed all of our data collection in a quiet office space with only ambient noise in the environment. It is important in our study that no additional background noise is present that could compromise the recordings.

**Speech Dataset:** For our experiments, we utilized speech samples from the TIDIGITS dataset [21]. This dataset contains audio samples of single digits (0-9) spoken by 10 speakers (5 female, 5 male). We utilize an audio sample of each digit, from each speaker, for a total of 50 different speech samples. Five samples were collected for each speaker (10), digit (10), and speech SPL (3) level resulting in a total of 1500 audio recordings recovered from the Echo Dot.

Additionally, we collected data where the speech played was a song (lyrics + music). Specifically, we used audio of the songs "Smooth" by Santana and "Blinding Lights" by The Weeknd. For each song, we cut out a 5 second portion of the beginning, middle, and end to use in our experiments which resulted in 6 different song clips. These samples will be used to attempt song recognition with the Shazam application, which uses a novel and sophisticated algorithm that can identify a song from a small snippet of audio.

**Collection Steps:** For each instance of data collection, we followed the same set of steps to generate and retrieve the Echo Dot recorded samples. We begin with the researcher manually activating the Echo Dot by issuing the wake word, "Alexa". Once the smart speaker has been awakened and is actively listening for the user's command (indicated by light ring glowing blue), we press the button on our PJD to manually start the noise injection. After the PJD has been started, the normal speech sample is played from the nearby Bluetooth speaker. After the speech sample finishes playing, we allow the injected noise to continue until the Echo Dot has finished its recording (indicated by the light ring powering down). Only after the recording has stopped do we manually stop our PJD from injecting noise. Once the audio has been recorded by the Echo Dot, the researcher accesses and saves the recordings from the Alexa Voice History web interface. Each recording made during data collection is saved as a .wav file and stored for later processing.

# 5 ATTACK DESIGN

## 5.1 Signal Processing of Recovered Audio

After data collection is complete, we amassed a set of 1500 recovered audio samples from the Echo Dot. In the interest of speech recognition tasks, we perform post-processing on the recovered audio signals to obtain the greatest results. We began the post-processing phase by manually trimming each recovered audio sample to about 1 second in length (enough to encompass the spoken digit).

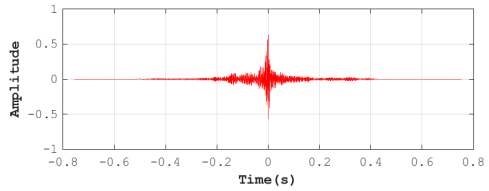
Next, we attempt to improve the quality of the normal speech in the recovered audio by applying a speech enhancement and noise reduction routine. We consider four such routines from the Matlab signal processing toolbox, VOICEBOX [17]. We performed some initial tests using the *specsub*, *spendred*, *ssubmmse*, and *ssubmmsev*. Each of these routines performs speech enhancement via some method including spectral subtraction, dereverberation, and minimum-mean square error (MMSE) with and without voice activity detection (VAD). Our initial tests found that the *ssubmmse* routine performed best in terms of speech enhancement and white noise reduction. Therefore, it was chosen as the noise filtering routine for our signal processing phase of the work.

## 5.2 Feature Extraction

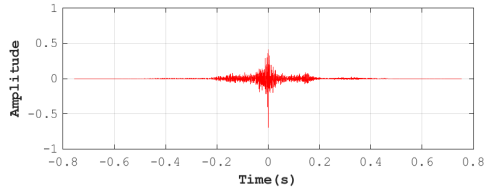
Once all audio samples in the dataset were processed, we perform MFCC feature extraction on each sample. MFCC features were chosen because they are widely used when attempting speech recognition tasks, especially when identifying spoken digits. For each audio sample we calculate the 13 MFCC coefficients which produces an Nx13 matrix where each column contains the values for each coefficient. From these coefficients we also calculate a single mean, minimum, maximum, and standard deviation value for each of the 13 MFCC coefficients. Additionally, we generate the first and second order differential coefficient values (e.g., Delta, log energy). Combining all of these values results in a 144-feature vector.

In addition to the full set of extracted features, we also use an attribute selection tool in Weka [10] to generate a filtered set of the

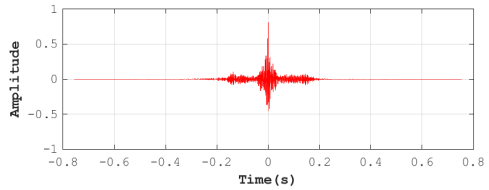
most important (e.g., most highly correlated) features in an attempt to achieve the greatest classification accuracies. We select the ClassifierAttributeEval class of evaluator (selecting the RandomForest classifier) and specify the BestFirst search algorithm. This produced a small set of the 17 most significant features which we also test in our classification models.




(a) 60 dB Speech




(b) 65 dB Speech




(c) 70 dB Speech

**Figure 4: Cross-correlation graphs comparing post-processed, Alexa recorded samples of speaker FAC saying the digit "One", for each speech SPL tested (60, 65, 70 dB), with the original raw audio file. The peaks found at lag=0 in these graphs indicate a strong correlation between our recovered signals and the original signal, further demonstrating the potential for an eavesdropping attack.**

### 5.3 Machine Learning Classification

We attempt different speech classification tasks using the dataset of VA recordings that we collected. Specifically, we explore *speech (digit) recognition*, *speaker identification*, and *gender identification*. In this section we will describe the performance results of our learning models that were trained for these speech recognition tasks. In our initial classification attempts we tested NaiveBayes, BayesNet, Logistic, MultiClass, and RandomForest classifier models. And for each of these classifiers we tested an 80:20 and 90:10 training/test data split, as well as 10-fold cross validation. We observe that across all classification tasks, the RandomForest classifier achieved the highest accuracies. Therefore, we highlight and report on the accuracies of the RandomForest classifier in the following sections.

## 6 SIGNAL ANALYSIS

### 6.1 Time & Frequency Spectrum

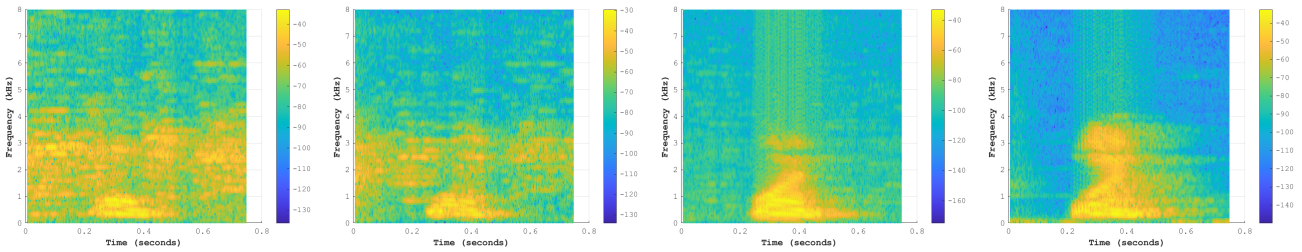
As an initial look at the post-processed recovered samples, we perform both time and frequency domain analysis which reveals the clear presence of normal speech in the recovered audio samples. Comparing the time domain graphs of samples from the different speech SPL settings, we find that the normal speech signal is visible in all of them. Figure 3 shows the time domain graphs generated from samples of the speaker FAC saying the digit "One" in each of the speech SPL settings. Although the success of noise cancellation varied across the different loudness levels of speech, being most successful when the normal speech was at its loudest, these time domain graphs confirm that the speech signal is maintained. We also find that the 70 dB speech sample collected with injected noise (Figure 3c) achieves a similar quality and strength of speech signal, after signal processing, to what we see in the clean sample collected without any injected noise. This suggests louder speech signals (70+ dB) may be too strong for even injected noise from a PJD to fully mask and protect from an eavesdropping attack. And even though the isolation of the speech signal is not as successful at lower speech SPLs, more sophisticated signal processing techniques could be used to obtain better results.

Our frequency spectrum analysis yielded similar observations that support what was seen in the time domain. In the spectrum graphs we can see the speech related frequencies are strong and present in all SPL settings. Additionally, we also see the presence of noise decrease as the source speech SPL was increased (e.g., noise cancellation improves). Figure 5 shows the spectrum graphs generated for the same samples from speaker FAC. We notice that in all cases the speech frequencies seem to be well identified and maintained through the post-processing for noise removal. The observations made in both the time and frequency domains are positive indicators that PJDs using white noise can be ineffective in the face of signal processing techniques, and therefore smart speakers equipped with a PJD could still be vulnerable to eavesdropping attacks. If we compare the 70 dB speech samples collected with injected noise (noisy, Figure 5c) and without (clean, 5d), we see the noisy sample can maintain almost all the same frequencies after signal processing, and at the same strength, as the clean sample.

### 6.2 Cross-Correlation

Continuing our analysis of the recovered samples, we perform normalized cross-correlation to compare post-processed samples from each SPL setting to the original raw audio used in our experiments. This allows us to gauge how much of the original signal was recovered after the post-processing routines were applied. Before determining the cross-correlation value, the two signals were aligned and a bandpass filter was applied to isolate frequencies between 150-1000 Hz (what appear to be the frequencies most related to the original speech).

Because the signals are aligned, we should find a peak at lag=0 in the cross-correlation graphs if the two signals are highly correlated. Figure 4 shows the cross-correlation graphs generated from samples of the speaker FAC saying the digit "One" in each of the speech SPL settings. Looking at the absolute amplitude in the cross-correlation graph, we find the normalized cross-correlation values for the 60, 65, and 70 dB samples are 0.63, 0.69, and 0.81 respectively. These values confirm a decent level of correlation between the signals (e.g., a significant amount of the original signal was recovered).




(a) 60 dB Speech (Noisy) (b) 65 dB Speech (Noisy) (c) 70 dB Speech (Noisy) (d) 70 dB Speech (Clean)

**Figure 5: Frequency spectrum graphs generated from post-processed, Alexa recorded samples (noisy) of speaker FAC saying the digit "One", for each speech SPL tested (60, 65, 70 dB); as well as a graph generated from a clean sample (no injected noise) with 70 dB speech. While normal speech related frequencies are present, we can see noise cancellation becomes more successful as the speech volume increases.**

**Table 1: Averaged results from PESQ and SNR analysis for each individual speaker.**

| Speaker ID | Alexa Recovered Audio | | | | | | Baseline | |
|---|---|---|---|---|---|---|---|---|
| | 60 dB | | 65 dB | | 70 dB | | | |
| | PESQ | SNR | PESQ | SNR | PESQ | SNR | PESQ | SNR |
| FAC | 1.2 | 8.8 dB | 1.3 | 8.6 dB | 1.7 | 7.7 dB | 1.8 | 15.7 dB |
| FBH | 1.2 | 9.8 dB | 1.3 | 10.6 dB | 1.3 | 10.2 dB | 1.6 | 14.9 dB |
| FCA | 1.2 | 10.2 dB | 1.2 | 9.8 dB | 1.3 | 9.2 dB | 1.4 | 20.7 dB |
| FDC | 1.2 | 7.1 dB | 1.2 | 7.5 dB | 1.2 | 6.7 dB | 1.5 | 24.0 dB |
| FEA | 1.2 | 8.6 dB | 1.3 | 9.0 dB | 1.5 | 9.3 dB | 1.6 | 16.8 dB |
| MAE | 1.2 | 8.2 dB | 1.2 | 8.0 dB | 1.5 | 6.5 dB | 1.6 | 23.5 dB |
| MBD | 1.2 | 8.8 dB | 1.3 | 9.0 dB | 1.4 | 7.1 dB | 1.7 | 22.1 dB |
| MCB | 1.3 | 8.0 dB | 1.3 | 8.2 dB | 1.3 | 7.5 dB | 1.7 | 17.9 dB |
| MDL | 1.3 | 8.7 dB | 1.1 | 9.6 dB | 1.5 | 7.9 dB | 1.8 | 16.2 dB |
| MEH | 1.2 | 9.3 dB | 1.1 | 10.1 dB | 1.4 | 6.6 dB | 1.7 | 16.2 dB |

**Table 2: Averaged results from PESQ and SNR analysis for both speaker genders.**

| Speaker Gender | Alexa Recovered Audio | | | | | | Baseline | |
|---|---|---|---|---|---|---|---|---|
| | 60 dB | | 65 dB | | 70 dB | | | |
| | PESQ | SNR | PESQ | SNR | PESQ | SNR | PESQ | SNR |
| Male | 1.2 | 8.8 dB | 1.3 | 9.0 dB | 1.4 | 7.9 dB | 1.6 | 18.4 dB |
| Female | 1.2 | 8.6 dB | 1.2 | 9.0 dB | 1.4 | 7.1 dB | 1.7 | 19.2 dB |

## 6.3 SNR & PESQ

Another way that we evaluated our recovered samples was using metrics that describe speech presence and quality in noisy audio. Specifically, we look at Signal-to-Noise Ratio (SNR) and Perceptual Evaluation of Speech Quality (PESQ) scores. SNR shows us how successful the post-processing was at reducing the noise and enhancing the speech frequencies. Further, the PESQ score rates the quality of the speech in terms of how perceptible it is to human listening. We believe these two standard metrics are useful for demonstrating the potential vulnerability of these audio samples to eavesdropping attacks. To highlight the greatest potential for speech to be recovered from noisy samples, we filtered our dataset for the SNR and PESQ analysis to include the samples from each experimental scenario that displayed the greatest speech leakage. For each speaker, digit, and speech SPL that we tested, we collected 5 different samples to build our 1500 sample dataset. Therefore, choosing the best sample out of 5 for each scenario resulted in a filtered dataset of 300 samples that we used to generate average SNR and PESQ scores for each speaker and gender. Table 1 and Table 2 show these average scores summarized per speaker and gender, respectively.

**Signal-to-Noise Ratio:** We calculated the SNR values of each individual audio sample using the equation:

$$SNR = 20 * log10(\frac{std(signal)}{std(noise)})$$

The variable *signal* refers to the recovered signal from each scenario, and the variable *noise* is the raw noise signal that was injected during the scenarios. The values produced by this equation were verified using the snr() function built into Matlab. We found the average SNR values for all speakers, in each experimental scenario, was positive. This was also seen when the values were averaged per

**Table 3: Summarized classification results (using Random Forest) observed for the different speech tasks, feature sets, and speech SPLs considered in our experiments. Bold accuracies indicate the highest values observed for each speech task and feature set.**

| Classification Task | # Classes | # Features | Speech SPL (dB) | Classification Accuracy (%) | | |
|---|---|---|---|---|---|---|
| | | | | 80:20 | 90:10 | 10-Fold CV |
| Speech (Digit) Recognition | 10 (digits, 0-9) Random Guess: **10%** | 144 (ALL) | 60 | 30 | 22 | 29.8 |
| | | | 65 | **40** | 34 | 37.2 |
| | | | 70 | 36 | 36 | 39.6 |
| | | 17 (filtered) | 60 | 24 | 22 | 28.4 |
| | | | 65 | 34 | **46** | 36.6 |
| | | | 70 | 35 | 42 | 36 |
| Speaker Identification | 10 (speakers) Random Guess: **10%** | 144 (ALL) | 60 | 30 | 40 | 38.4 |
| | | | 65 | **51** | 50 | 46.2 |
| | | | 70 | 39 | 38 | 39.4 |
| | | 17 (filtered) | 60 | 32 | 36 | 40.2 |
| | | | 65 | 49 | **50** | 43.4 |
| | | | 70 | 37 | 42 | 40.6 |
| Gender Identification | 2 (Male, Female) Random Guess: **50%** | 144 (ALL) | 60 | 76 | 69 | 75.6 |
| | | | 65 | **80** | **80** | 76.8 |
| | | | 70 | 66 | 76 | 70 |
| | | 17 (filtered) | 60 | 77 | 71 | 76.9 |
| | | | 65 | 76 | **78** | 74.8 |
| | | | 70 | 71 | 74 | 69.2 |

gender. Although the SNR scores do not reveal any particular pattern or trend across the different SPL settings (i.e., similar amount of noise can be removed at all speech levels), it is important to recognize that all recovered and processed samples produced positive SNR values, with most scenarios averaging an SNR of 8 dB or higher. These positive values indicate the recovered speech signal is greater than the noise remaining in the sample post-processing. We also collected baseline samples of the normal speech audio (without any injected noise) which produced the highest SNR values. The decrease in SNR of our recovered samples, compared to the baseline, reflects noise that still remains even after signal processing.

**Perceptual Evaluation of Speech Quality:** PESQ is the metric that is most related to human perceived intelligibility of the recovered samples [29]. The results from our PESQ analysis support our previous observations of speech leakage. To calculate PESQ scores for individual audio samples, we used a python PESQ wrapper [35]. The provided function takes the recovered audio sample and the original raw audio sample as input and compares them to determine the quality of speech contained in the sample. After averaging the scores for both genders and for each speaker, we find positive scores above 1.0 for all scenarios. Additionally, we observed a consistent and expected pattern among the averaged values where the PESQ scores increase (albeit slightly) as the original speech SPL increases. Further, we also find that the PESQ scores calculated are comparable to the scores of the baseline samples. This means that the speech content we were able to recover has a similar intelligibility, in terms of human perception, to the raw recordings we collected without any injected noise. PESQ is a significant indicator of potential speech leakage, and the results for our samples suggest the quality of speech that can be acquired via our attack model (to achieve speech recognition) is on par with an actual microphone.

## 7 CLASSIFICATION RESULTS

In this section we will discuss the accuracies we achieved for speech (digit) recognition, speaker identification, and gender identification tasks. We attempt each of these tasks under a few different parameter settings. Table 3 summarizes the classification accuracies observed for each speech task, feature set, and speech loudness (SPL) that we consider.

**Digit Recognition:** For digit recognition, 40% classification accuracy was achieved when the model was trained on the full set of 144 features and an 80:20 split. When the feature set was filtered to the 17 most significant features, we observe an increased maximum accuracy of 46% for the 90:10 split. Interestingly, both of these accuracies were observed for the 65 dB speech data. Although 46% may not sound impressive, if we consider that random guessing has a 10% classification accuracy (for classifying the 10 digits), we see that the accuracies observed in our experiments are an improvement. The 10-fold Cross Validation results confirm that the classification accuracy increases as the source speech loudness is increased past 60 dB. Interestingly, the results show that the 65 dB and 70 dB source speech have similar classification results, with the 65 dB speech achieving a slightly better classification accuracy with the filtered feature set. This may suggest the Alexa noise cancellation that is applied is similarly effective for both speech SPLs, and that there is no significant benefit for the loudest speech level (70 dB).

**Speaker Identification:** The speaker identification accuracies that we see are also an improvement on the 10% accuracy of random guessing (for classifying the 10 speaker). When we trained our model on the full set of features we see a maximum classification accuracy of 51% for the 80:20 split. Similarly, we observed a maximum accuracy of 50% for the 90:10 split when the model was trained on the filtered set of features. 10-fold Cross Validation revealed a

different pattern of success for speaker identification compared to what we observed for digit recognition. The results show that the 65 dB source speech had the best classification results for both feature sets. This indicates that Alexa noise cancellation is able to preserve more speech frequencies (e.g., those needed for speaker identification) when the speech is around 65 dB. It's possible that louder volume speech may produce frequencies at power levels that can be confused with that of noise, and are therefore removed during noise cancellation (inline with observations from digit recognition).

**Gender Identification:** We observe decent classification accuracies for the gender recognition task for both feature sets. When we trained on the full set of features, we achieved 80% classification accuracy for both train/test data splits. And when we used the filtered set of features, we observed a maximum accuracy of 78% for the 90:10 data split. This is a significant improvement to the 50% accuracy achieved through random guessing (classifying two genders), indicating the clear potential for successful gender recognition by an attacker. 10-fold Cross Validation for the gender identification task revealed the best classification results were achieved when speech was at the lower volumes (60, 65 dB). Similar to our previous observations, these results suggest that gender-specific information is better preserved at volumes lower than the maximum we tested (70 dB). Again, at the louder volume level the power of the speech frequencies may be confused with noise and get removed. And a 5%+ decrease in classification accuracy (for both feature sets) when the speech was at 70 dB suggests important speech information required for *gender identification* is lost at that volume.

**Song Recognition:** Lastly, we conducted a small evaluation of song recognition potential using the recovered and processed samples from Alexa Voice History. We played two songs and isolated short clips (5 seconds) from the beginning, middle, and end of each song. These samples consisted of a mix of music with and without spoken lyrics. To evaluate these samples we used the Shazam song recognition app which employs an efficient, scalable, noise and distortion resistant song identification algorithm [44]. We found that all song samples that were prepared could be successfully identified. Each sample was tested five times and we observed 100% song recognition accuracy for all samples. Song recognition can be considered one of the easier tasks to accomplish that can still reveal sensitive information. An attacker could use already available tools, or even design their own customized song recognition algorithm. The algorithm used for our work performs a combinatorially hashed time-frequency analysis of the audio to recognize a song with very small samples of the original audio. Further, songs can be easily identified with snippets of either instrumental music or lyrics being sung. In regards to privacy, song recognition can reveal the user's unique and personal interests.

## 8 DISCUSSION

**Increased Potential for Compromising Speech:** Through our experiments, we determined that a RandomForest classifier can be used to improve the potential for certain speech recognition tasks (beyond the accuracies obtained from random guessing). Our results demonstrate the clear potential for an attacker to compromise user speech, to some extent, even when under the protection of a PJD. We found that full speech (digit) recognition and speaker identification are more challenging, while successful gender recognition seems more likely in a real-world attack scenario. Further, an attacker with more extensive knowledge of signal processing, or with improved techniques, could achieve even better classification accuracies.

**Speech SPL Observation:** The classification results we observed for each speech task and feature set revealed an interesting pattern. In all scenarios we find that the maximum accuracies were achieved using source speech at 65 dB. This is unexpected as we would think that the 70 dB source speech data would produce the greatest results for speech classification tasks. Through the process of collecting the Alexa voice history recordings we noticed that some noise cancellation is already applied. However, the success of this noise cancellation was not consistent and resulted in samples recorded under the same settings having different levels of noise still present. Therefore, we speculate that one reason for the improved classification success of the 65 dB samples could be better noise cancellation performance on user speech that is closer to 65 dB. So noise cancellation may be less effective on user speech that is 70 dB because it is furthest from the range of normal human speech. Speech related features could be filtered because the increased SPL (power) of the speech signal is mistaken as noise.

## 9 CONCLUSION

In this work we look to explore the effectiveness of Protection Jamming Devices (PJD) that use GWN for masking user speech from eavesdropping attacks. These devices are used to continuously inject a masking sound into the microphones of VA smart speakers in order to block the device from accepting the user's commands. An assumption is that this hinders any potential VA speaker eavesdropping attacks. However, with current signal processing techniques (i.e., noise reduction, speech enhancement) there exists a potential for the user's speech to be compromised by an attacker that can access smart speaker recordings. Through a process of data collection, post-processing, feature extraction, and model training, we were able to demonstrate greater classification accuracies (than random guessing) of 46%, 51%, and 80% for speech (digit) recognition, speaker identification, and gender identification, respectively.

**Future Work:** As this work provides a first exploration of PJDs (using GWN) effectiveness for masking speech recorded by a smart speaker device, we could not feasibly test all possible experimental parameters. Therefore, a few key directions remain that we can explore in the future to further develop this work (and our overall understanding of PJD device effectiveness). First, we decided on a Gaussian white noise for our jamming signal for its popular use in other speech masking solutions. Aside from GWN, there are other jamming signal types that we would like to explore including chatter noise and ultrasonic sound. These signal types can provide more diversity in the jamming noise or even affect the performance of the microphone mechanics. Next, we would like to test other PJD configurations such as using 6 tiny speakers to inject noise in the microphone array (like Home Wave), or encasing all of the components in a housing shell (like Project Alias). Lastly, we would like to explore new or more extensive signal processing techniques to improve the noise cancellation and speech enhancement.

## ACKNOWLEDGMENTS

We would like to give special thanks to the set of anonymous reviewers for their valuable feedback on this paper. This work is partially supported by the National Science Foundation (NSF) under the grants: CNS-1714807, CNS-2030501, CNS-2139358.

## REFERENCES

[1] Speech Masking 2016. *Our Technology.* Speech Masking. https://www.speechmasking.com/Technology

[2] 2020. *Data security and privacy on devices that work with Assistant.* https://support.google.com/googlenest/answer/7072285?hl=en

[3] Amazon 2020. *Echo Dot (3rd Gen) - Smart speaker with Alexa.* Amazon. https://www.amazon.com/Echo-Dot/dp/B07FZ8S74R

[4] Google 2020. *Google Home Mini.* Google. https://store.google.com/us/config/google$_{h}$ome$_{m}$ini

[5] Consumer Watchdog 2020. *How Google and Amazon are 'spying' on you.* Consumer Watchdog. https://www.consumerwatchdog.org/privacy-technology/how-google-and-amazon-are-spying-you

[6] 2020. *Is Alexa Recording?* https://www.amazon.com/is-alexa-recording-conversations/b?ie=UTF8&node=21219697011

[7] NPR 2020. *NPR and Edison Research Report.* NPR. https://www.npr.org/about-npr/794588984/npr-and-edison-research-report-60m-u-s-adults-18-own-a-smart-speaker

[8] 2020. *Paranoid Home Devices - Home Wave.* https://paranoid.com/products

[9] Sound Management Group 2020. *Sound Masking for Offices.* Sound Management Group. https://soundmanagementgroup.com/products/sound-masking/

[10] University of Waikato 2020. *WEKA: The workbench for machine learning.* University of Waikato. https://www.cs.waikato.ac.nz/ml/weka/index.html

[11] Pro Acoustics 2021. *Sound Masking Systems.* Pro Acoustics. https://www.proacousticsusa.com/complete-sound-systems/commercial-sound-systems/sound-masking-systems.html

[12] Alina Bradford. 2020. *Paranoid prevents smart speakers from eavesdropping on you.* Digital Trends. https://www.digitaltrends.com/home/paranoid-prevents-smart-speakers-from-eavesdropping/#:~:text=Paranoid%20is%20a%20device%20that,%E2%80%9CParanoid%E2%80%9D%20before%20each%20command.

[13] April White. 2018. *A Brief History of Surveillance in America.* Smithsonian Magazine. https://www.smithsonianmag.com/history/brief-history-surveillance-america-180968399/

[14] Bjorn Karmann. 2018. *Project Alias.* https://bjoernkarmann.dk/project$_{a}$lias

[15] Bjorn Karmann. 2018. *Project Alias.* Instructables Circuits. https://www.instructables.com/id/Project-Alias/

[16] Brittany Vance. 2020. *New device stops your smart speaker from listening without a safe word.* The American Genius. https://theamericangenius.com/gadgets/new-device-stops-your-smart-speaker-from-listening-without-a-safe-word/

[17] Mike Brookes. 2017. *VOICEBOX: Speech Processing Toolbox for MATLAB.* http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[18] Y. Chen, Huiying Li, Shan-Yuan Teng, S. Nagels, Zhijing Li, Pedro Lopes, B. Zhao, and H. Zheng. 2020. Wearable Microphone Jamming. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020).

[19] P. Cheng, I. E. Bagci, J. Yan, and Utz Roedig. 2018. Towards Reactive Acoustic Jamming for Personal Voice Assistants. *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security* (2018).

[20] Chris Matyszczyk. 2020. *This weird new gadget stops Amazon's Alexa spying on you.* ZDNet. https://www.zdnet.com/article/this-weird-new-gadget-stops-amazons-alexa-spying-on-you/

[21] Dan Ellis. 2003. *Clean Digits.* Columbia University. https://www.ee.columbia.edu/~dpwe/sounds/tidigits/

[22] Danny Bradbury. 2020. *Smart speakers mistakenly eavesdrop up to 19 times a day.* https://nakedsecurity.sophos.com/2020/02/25/smart-speakers-mistakenly-eavesdrop-up-to-19-times-a-day/

[23] Don Sweeney. 2019. *Here's how to stop Amazon employees from eavesdropping on your Alexa conversations.* https://www.sacbee.com/news/nation-world/national/article229121874.html

[24] Eric Limer. 2019. *This Brilliant Home Assistant Add-On Makes Eavesdropping Completely Impossible.* Popular Mechanics. https://www.popularmechanics.com/technology/security/a25894138/amazon-echo-google-home-project-alias-white-noise-eavesdropping-defense/

[25] J. Fingas. 2019. *Florida police obtain Alexa recordings in murder investigation.* engadget. https://www.engadget.com/2019-11-02-florida-police-obtain-alexa-recordings-in-murder-case.html

[26] Mark D. Fletcher, Sian Lloyd Jones, Paul R. White, Craig N. Dolder, Timothy G. Leighton, and Benjamin Lineton. 2018. Effects of very high-frequency sound and ultrasound on humans. Part I: Adverse symptoms after exposure to audible very-high frequency sound. *The Journal of the Acoustical Society of America* 144, 4 (2018), 2511–2520. https://doi.org/10.1121/1.5063819

[27] Sidney Fussell. 2020. *Meet the Star Witness: Your Smart Speaker.* Wired. https://www.wired.com/story/star-witness-your-smart-speaker/

[28] Geoffrey A. Fowler. 2020. *Alexa has been eavesdropping on you this whole time.* The Washington Post. https://www.consumerwatchdog.org/privacy-technology/how-google-and-amazon-are-spying-you

[29] Mordechai Guri, Yosef Solewicz, Andrey Daidakulov, and Yuval Elovici. 2017. SPEAKE(a)R: Turn Speakers to Microphones for Fun and Profit. In *Proceedings of the 11th USENIX Conference on Offensive Technologies (WOOT'17).* USENIX Association, USA, 13.

[30] Jason Cohen. 2020. *Smart Speaker Sales Soar as Owners Buy Multiple Devices.* https://www.pcmag.com/news/smart-speaker-sales-soar-as-owners-buy-multiple-devices

[31] Jeffrey Lang. 2020. *Yes, your smart devices are spying on you. Which ones are the worst?* https://movietvtechgeeks.com/yes-your-smart-devices-are-spying-on-you-which-ones-are-the-worst/

[32] Yeongseok Kim and Youngjin Park. 2017. Effect of active noise control and masking sound on speech intelligibility. *Applied Acoustics* 123 (2017), 152–157. https://doi.org/10.1016/j.apacoust.2017.02.021

[33] Alexey Krasnov, Edward R Green, Bret Engels, and Barry Corden. 2019. Enhanced speech privacy in office spaces. *Building Acoustics* 26, 1 (2019), 57–66. https://doi.org/10.1177/1351010X18798105

[34] L. Li, Manni Liu, Yuguang Yao, Fan Dang, Zhichao Cao, and Y. Liu. 2020. Patronus: Preventing Unauthorized Speech Recordings with Support for Selective Unscrambling. *Proceedings of the 18th Conference on Embedded Networked Sensor Systems* (2020).

[35] ludlows. 2019. *PESQ (Perceptual Evaluation of Speech Quality) Wrapper for Python Users (narrow band and wide band).* GitHub. https://github.com/ludlows/python-pesq

[36] Mark Wilson. 2019. *This is the first truly great Amazon Alexa and Google Home hack.* Fast Company. https://www.fastcompany.com/90290703/this-is-the-first-truly-great-amazon-alexa-and-google-home-hack?partner=rss&utm$_{s}$ource=feedburner&utm$_{m}$edium=feed&utm$_{c}$ampaign=feedburner+fastcompany&utm$_{c}$ontent=feedburner

[37] C. Phunruangsakao, P. Kraikhun, S. Duangpummet, J. Karnjana, M. Unoki, and W. Kongprawechnon. 2020. Speech Privacy Protection based on Controlling Estimated Speech Transmission Index. In *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON).* 628–631.

[38] Shannon Liao. 2019. *This project hacks Amazon Echo and Google Home to protect your privacy.* The Verge. https://www.theverge.com/circuitbreaker/2019/1/15/18182214/amazon-echo-google-home-privacy-protection-project-white-noise

[39] Bożena Smagowska and Małgorzata Pawlaczyk-Łuszczyńska. 2013. Effects of Ultrasonic Noise on the Human Body—A Bibliographic Review. *International Journal of Occupational Safety and Ergonomics (JOSE)* 19, 2 (2013), 195–202. https://doi.org/10.1080/10803548.2013.11076978

[40] Spencer Ackerman. 2012. *CIA Chief: We'll Spy on You Through Your Dishwasher.* Wired. https://www.wired.com/2012/03/petraeus-tv-remote/

[41] Ken Sun, Chen Chen, and Xinyu Zhang. 2020. "Alexa, Stop Spying on Me!": Speech Privacy Protection Against Voice Assistants.

[42] Harshavardhan Sundar, Weiran Wang, Ming Sun, and Chao Wang. 2020. Raw Waveform Based End-to-end Deep Convolutional Network for Spatial Localization of Multiple Acoustic Sources. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'20).* 4642–4646.

[43] Trevor Timm. 2016. *The government just admitted it will use smart home devices for spying.* The Guardian. https://www.theguardian.com/commentisfree/2016/feb/09/internet-of-things-smart-devices-spying-surveillance-us-government

[44] Avery Wang. 2003. An Industrial Strength Audio Search Algorithm. In *4th International Conference on Music Information Retrieval (ISMIR'03).*

[45] Zack Whittaker. 2018. *Judge orders Amazon to turn over Echo recordings in double murder case.* TechCrunch. https://techcrunch.com/2018/11/14/amazon-echo-recordings-judge-murder-case/

[46] Minhua Wu, Kenichi Kumatani, Shiva Sundaram, Nikko Ström, and Björn Hoffmeister. 2019. Frequency Domain Multi-channel Acoustic Modeling for Distant Speech Recognition. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19).* 6640–6644.

[47] Jun Yang. 2018. Multilayer Adaptation Based Complex Echo Cancellation and Voice Enhancement. 2131–2135.