Speaker Misactivations: Say something if you cannot hear me

Matthew Wong
University of Maryland, College Park

Timothy Lin
University of Maryland, College Park

Geng Liu
University of Maryland, College Park

Abstract

As smart speakers become more common and normalized, there is the potential for sensitive information to be captured in the audio of always-on smart speakers; smart speakers can be privy to very sensitive information that may be sent over the network or stored in manufacturers' servers. We conducted a study on a set of smart speakers and tested the misactivation rate of the speakers using spoken word audio. In this study, we are not looking at potential attacks to retrieve this data, but at the quantity and likelihood of vulnerable information (transmitted over the internet or stored elsewhere not on the device). While there are still many areas to explore in this space, we note some novel observations about the behavior of these smart speakers towards misactivations and the discrepancy between the Amazon Echo Dot and Google Home Mini devices.

1 Introduction

Smart speakers, and their associated smart personal assistants, have become more and more popular in recent years in the home, where they are privy to private conversations. They are able to activate and respond to a user's command upon hearing a *trigger word* particular to each speaker or manufacturer (some manufacturers recently also allow customization of the trigger word). However, this always-on functionality means that the speaker must constantly be listening to its environment and may record and transmit audio not intended for the smart assistant across the internet. In the event of these *misactivations*, sensitive audio from the speaker's environment may be unintentionally recorded and subject to another attack in which an attacker gains access to many such recorded clips and transcripts.

Thus, it is important to understand how often and on what kinds of inputs these smart speakers activate. Some previous studies have looked at ways to find misactivations or generate their own misactivations. Our approach focuses on the words themselves and speaker accents, rather than clips of audio in which the immediate cause of the misactivation may be less obvious.

2 Related Works

There have been several previous works showing the feasibility of exploits on IoT devices using various side channels. Light Commands was a project where researchers used laser light to inject malicious commands into several voice-controlled devices such as smart speakers, tablets, and phones across large distances and through glass windows, causing them to misactivate [6]. They accomplished this attack using the fact that microphones convert sound waves into electrical signals. Since light is a wave as well, if a light wave can be generated such that the shape (frequency and amplitude) could match a certain sound wave, then the microphone can interpret the light beam as genuine audio, causing the device to misactivate.

The Moniotr Lab conducted a study in which the researchers played 134 hours of TV shows at two different locations twice (for a total of 536 hours) to some smart speakers trying to activate them without the use of their "key words" [1] (which we refer in this paper as "trigger words"). The analysis includes more than one million words of dialogue from different regions. They found that all the devices are misactivated around 0.4 times per hour on average. In all of those misactivations, audio was collected from the environment and sent over the internet, which could potentially cause a privacy leak.

Given audio data recorded from a nearby smart speaker, Zarandy et al. were able to sometimes reconstruct a password (or other information) typed into a smartphone [7]. They demonstrated the danger of leaked access to surreptitiously recorded audio in an attack on extremely sensitive data (namely, passwords) after gaining access to audio from a smart speaker. If not directly

listening on the device, this attack is more effective the more recorded audio is sent over the network and stored in corporate servers (which can be artificially increased by misactivations).

Schönherr et al. performed several experiments where they played hours of TV shows, news, and word datasets to a multitude of smart speakers in English, German, and Chinese and also changed the gender of the speaker to test for potential gender bias [5]. Furthermore, they constructed their own "automatic speech recognition system" to synthesize misactivation words. From their experiments, they have published about 1000 words that have misactivated the devices; however their misactivations dataset is currently not available to the public.

Our experiments differ from previous work in that we examine the accent of the speaker and the effect it has on the number of misactivations and isolating specific words that cause misactivations independent of background noise. We performed experiments where the audio used the English language but the accent of the speaker varied. Dubois et al. only played shows in the English language across two locations without varying the accent of the dialogue in the TV shows [1]. Even though the speakers could have had different accents in the shows, they did not perform an in-depth analysis on whether or not different accents affected the number of misactivations that occur. Schönherr et al. only performed experiments varying the language spoken in TV shows, news, and word datasets as well as varying the gender of the speaker [5]. Due to time constraints, our experiments only examined three of the dozens of available accents; however, we were able to make some interesting observations.

3 Methods

3.1 Measuring Misactivations

We look at 3 sources of data to determine when a smart speaker has misactivated:

Light activations. Usually, the light on a smart speaker activating is an indication that the speaker is actively listening for a request and may transmit audio.

Network activity. When a speaker detects and responds to a request, there will be increased network activity to the server indicating that recorded audio is sent to the server and a response may be sent back in response.

Server records. Companies like Amazon and Google will also store records of when their speakers were activated, with metadata such as time of activation, request transcript (or recording), response to the user (as a transcript or recording), and other information, depending on the manufacturer and the specific details of the activation.

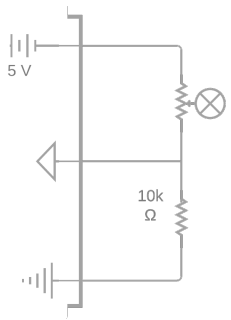

Figure 1: Circuit diagram for the Arduino and photoresistor voltage divider setup.

To measure when the light activates, we position a photoresistor in front of the smart speaker's light to measure the ambient light level (see Figure 1). We record the start and end time for each period during which the light level was above a certain threshold on an Arduino, and this stream of light activation and deactivation times is sent to the Raspberry Pi. In addition to collecting the light activation data from the Arduino, the Raspberry Pi (or similar device) is also set up as an access point for a wireless local network for the smart speaker to connect via; we ran Wireshark on this device to capture packets in order to later extract timestamps and packet sizes. We also manually fetched activation records from Google and Amazon's websites and manually cross-referenced their records with our expected activation times. A visual representation of this process is depicted in Figure 2.

3.2 Generating & Playing Potential Misactivation Audio

We generate several synthetic audio datasets in order to measure which specific words will cause misactivations. To generate a dataset, we pulled words/sentences from several datasets:

1. Top 10k most common English words, as collected by Google [2]
2. First 2 volumes of the public domain book *Clarissa* [4]
3. Google Natural Language Questions dataset [3]

(For (2) and (3), we trained an 8-gram model with Laplace smoothing and generated 2500 sentences as our "word" list.)

We then used Google Translate's text-to-speech engine to speak each of the words or sentences in

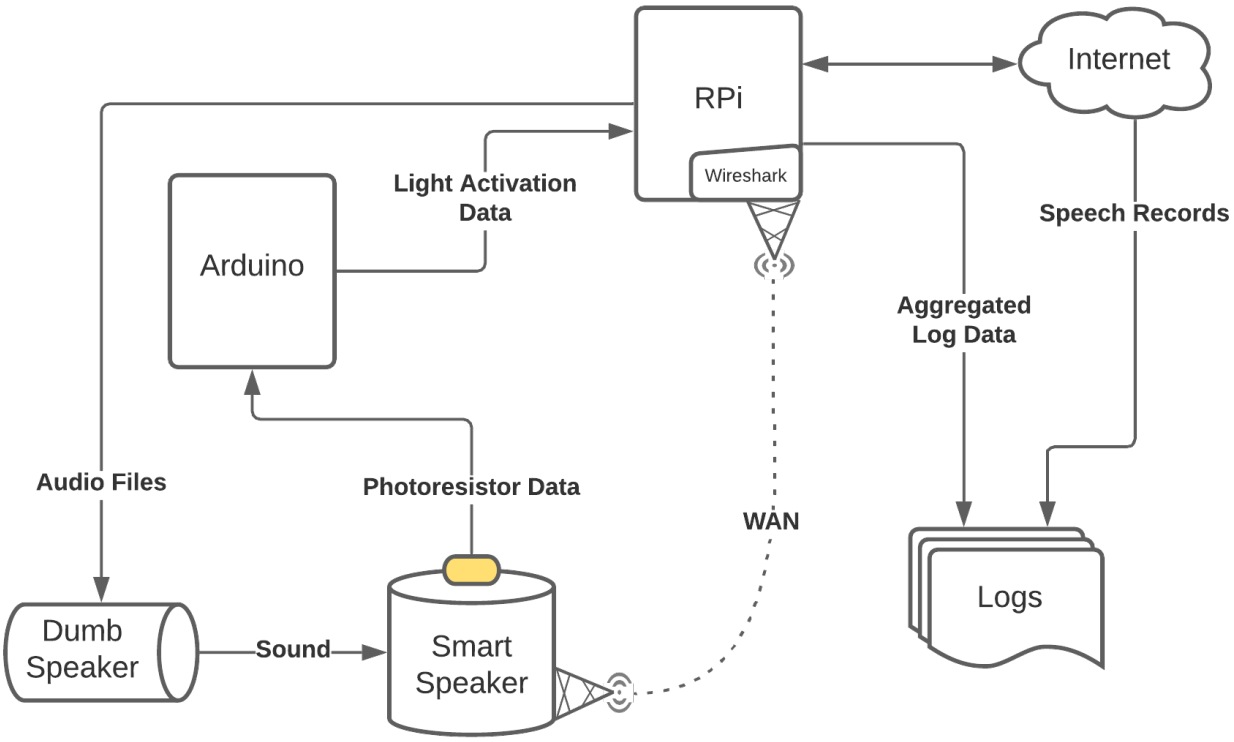

Figure 2: Flow of data within our hardware & network setup. A Raspberry Pi commands a dumb speaker to play audio at the smart speaker. An Arduino monitors for any changes in light levels and relays light activation timestamps to the Raspberry Pi; the Raspberry Pi also acts as a bridge and monitors for network traffic to and from the smart speaker. The network data collected using Wireshark and light sensor activation data are combined with the server records for analysis.

the default-gendered voice in each of three accents: US/American, UK/British, and Indian. Three trigger words were also interspersed in all of the datasets: "Ok Google," "Hey Alexa," and "Hey Siri" as a way to make sure that we could get normal activations using this method. One of these 3 trigger words was chosen uniformly at random to be spoken every 50 words spoken. In later experiments, we would also ask a question if an activation was detected immediately after any word; this checked if the speaker was actually ready to receive voice commands.

Immediately after playing each word or sentence, we wait for a small delay to give the speaker some time to activate in case. If the speaker light activates, then we pause playing new words until after the light turns off. This both gives a longer period of time to catch the speaker's activation and to prevent the next spoken word from causing the speaker/assistant to further respond.

4 Experiments

We performed several experiments with the datasets mentioned above on two IoT devices: the Amazon Echo Dot (3rd Generation) and the Google Home Mini (1st generation). For the top 10k most common English words dataset, we ran three experiments (one experiment for each accent mentioned above) using the methodology described in Section 3.2. In each experiment, we ran three trials for the Amazon Echo Dot and one trial for the Google Home Mini. Furthermore, we ran two experiments using the same methodology with the 5000 phrases generated from the 8-gram model using the *Clarissa* and Google Natural Language Questions datasets. However, we only ran one trial of both these experiments using the American accent to speak the sentences for both devices.

For example, for one "trial" of the 10k common words experiment for the Echo Dot, we can extract the data visualized in Figure 3. We are primarily interested in the activations (spikes), especially when not all 3 of our measurement indicators matched up.

5 Results and Analysis

5.1 Echo Dot

After examining the server logs of the Echo Dot, it seems as though whenever the Echo Dot activates on a word, there is some algorithm on the Amazon server that determines whether or not that word should be treated as a trigger word or not. We found that audio recordings were usually flagged as either "alexa", "hey alexa", or "Audio was not intended for Alexa". We will define a *full server activation* as a word where the recording on the server was flagged as "alexa" or "hey alexa" and the Echo Dot kept recording a few seconds after that word, waiting for a phrase from the user. If a phrase is said by the user, it will be recorded and flagged as the text of the recording or if silence is recorded, then the server will flag the recording as "No text stored". Furthermore, we will define a *partial server activation* as a word where the recording was flagged as "Audio was not intended for Alexa" and the Echo Dot immediately stopped recording after that word. We will mainly be looking for non-trigger words that have a full server activation. This misactivation type is the most dangerous since sensitive information could be recorded after the non-trigger word is said and sent to the server.

5.1.1 Top 10k Most Common English Words Dataset

All of the trigger and non-trigger words that caused light activations were recorded on the Amazon server. 12.68% (9 out of 71) of the non-trigger words that caused a light activation were also found to be full server activations (Figure 5). The rest of the non-trigger words that had light activations were found to be partial server activations. Moreover, 8.17% (45 out of 551) of the detected trigger words that had light activations were found to be partial server activations. The rest of the detected trigger words that had light activations were found to be full server activations. None of the undetected light activation trigger words had any recordings on the server.

For the non-trigger words that were full server activations, the average time of light activation was 2.92 seconds. On the other hand, the average light activation time for the non-trigger words that were partial server activations was 1.84 seconds. We can observe that the average light activation time for non-trigger words that were full server activations was almost one second longer than the average light activation time for non-trigger words that were partial server activations. It is also interesting to see that the average light activation time for detected trigger words that were full server activations was about 6.72 seconds. The average light activation duration for trigger words that were full server activations is almost double the average light activation duration for non-trigger words that were full server activations. Furthermore, the average light activation time for detected trigger words that were partial server activations was 1.83 seconds. This average time is very similar to the average time of a non-trigger word that was a partial server activation.

There are several conclusions we can draw from our observations. Whenever the Echo Dot lights up, the audio is recorded on the server. There is no evidence where the Echo Dot recorded a word without lighting up. Furthermore, non-trigger words have a chance to be inter-

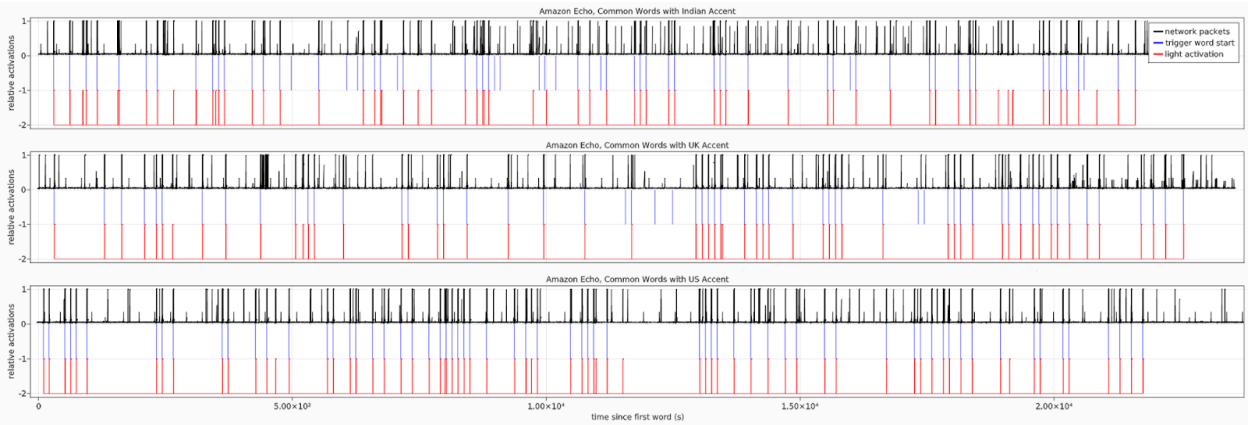

Figure 3: Experiment run data for the 10k common words dataset on the Amazon Echo Dot, for a single run of each of the 3 accents: Indian (top), UK/British (middle), and US/American (bottom). The blue lines represent when the trigger word for the device, "Hey Alexa," is spoken intentionally; the red lines represent times when the light is detected to be activated; and the black data shows relative network traffic on the local network.

Accent	Common Words	
	Sound Type	Example Words
American	*-x-*	Lexus, flexible, flexibility, Luxembourg
	co-	collection, collective, correction
British	*le-*	Lexmark, letter, lesser
	w-	workshop(s), webshots
	a-	Alexandria, Alexander, isa, Acer
Indian	*a-*	Alex, answer, Arthur, author, awesome
	co-	collection(s), collector(s), culture
	el-	electro, electricity, electric, elected

Table 1: Examples of words from the 10k most common words experiments that misactivated the Amazon Echo Dot.

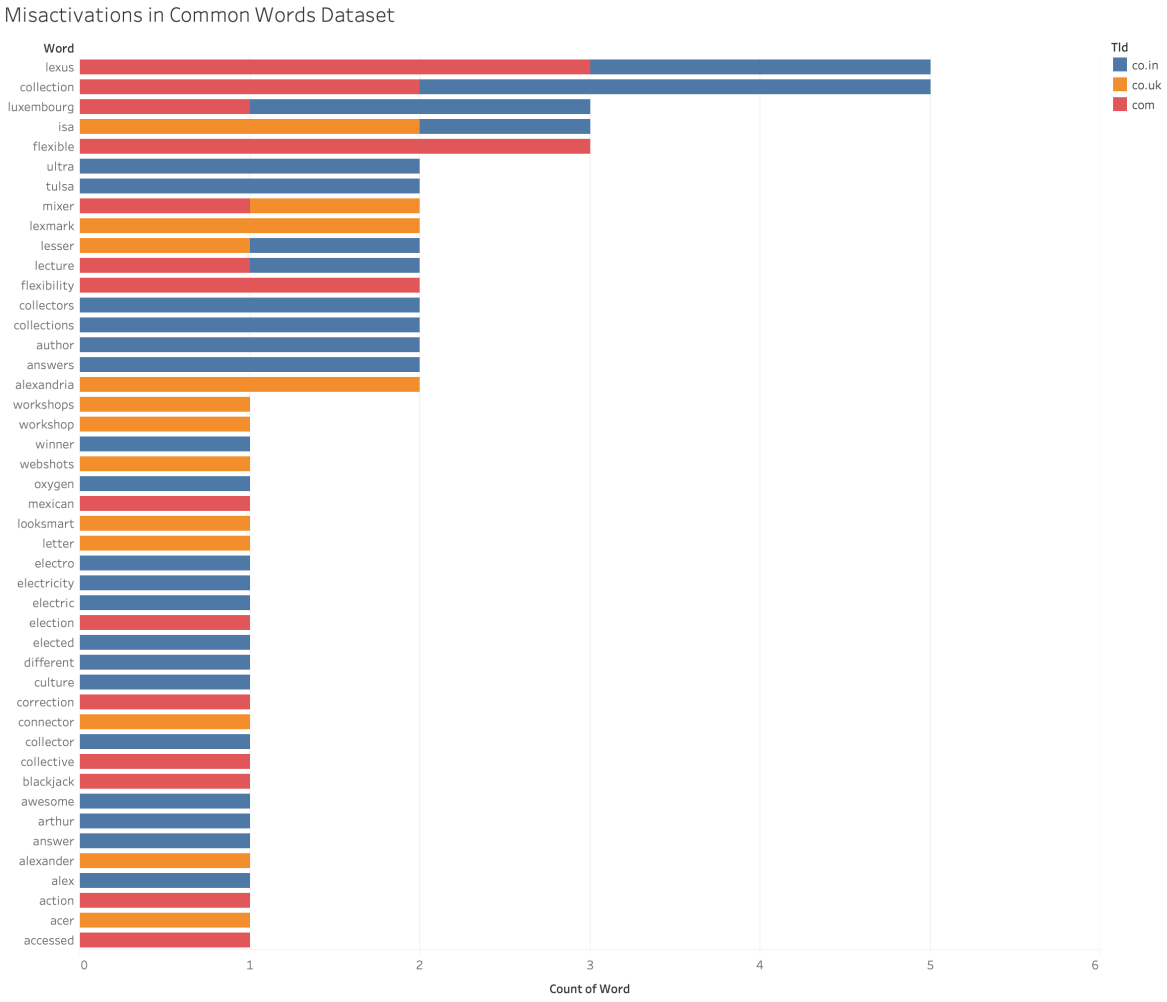

Figure 4: Count of all misactivations among all trials and experiments with the 10k most common English words dataset. There was a total of 71 non-distinct words that misactivated the Echo Dot among all trials and experiments. In the legend, "co.in" refers to experiments done with the Indian accent, "co.uk" refers to the experiments done with the British accent, and "com" refers to the experiments done with the American accent.

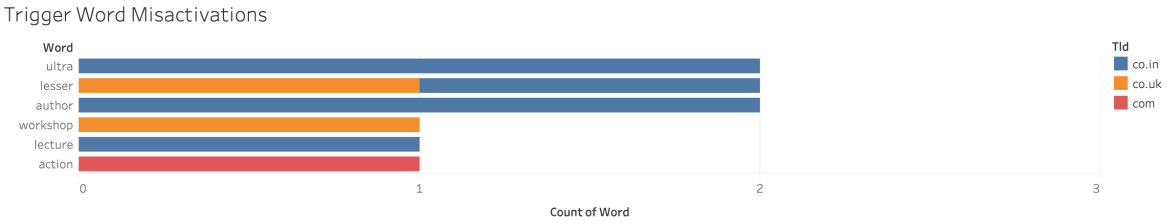

Figure 5: Count of misactivation words that acted as a normal trigger word ("full server activation") for the Echo Dot. There was a total of 9 non-distinct words that acted as a normal trigger word among all trials and experiments. In the legend, "co.in" refers to experiments done with the Indian accent, "co.uk" refers to the experiments done with the British accent, and "com" refers to the experiments done with the American accent.

preted as a trigger word with a full server activation, but the activation time for the non-trigger word is drastically less than the activation time for the trigger word, assuming both are full server activations. This is quite interesting since Amazon flags the non-trigger word as a trigger word, but the activation time is much less. In addition, it seems as though if the Echo Dot sent a word to the server, and the word was flagged as "Audio was not intended for Alexa", the activation times should be very similar no matter if the word was a trigger word or not. An interesting note is that the average time of activation when a non-trigger word is a full server activation is about half the average activation time of a trigger word that is a full server activation, but still greater than the average time of a non-trigger word that is a partial server activation.

From observing the count of all misactivation words (Figure 4), the British accent experiments had the least amount of misactivations with 16 misactivations, the American accent experiments had 20 misactivations, and the Indian accent experiments had the most with 35 misactivations. Furthermore, the Indian accent experiments had the most non-trigger words that were full server activations of 6, the British accent experiments had 2, and the American accent experiments had 1. For undetected trigger word activations, the American accent experiments only had 1, the British accent experiments had 17, and the Indian accent experiments had 30.

A lot of the misactivated words include some phoneme(s) of "Hey Alexa", confirming findings of previous research [1, 5]; however there were also a lot of recurring words that did not sound like "Hey Alexa". For example, "answers" and "author" activated the Echo Dot in two trials across the three trials for the Indian accent experiment. It is possible that these words could be recurring misactivation words when said in the Indian accent, but more trials would need to be done. In addition, different accents introduced new words that were not found with other accents. For example, *co-* words and *a-* words only misactivated the Echo dot when the words were said in the Indian accent and not in the American or British accent.

On a side note, we mentioned that we also found an occurrence where the Echo Dot recorded a request activating after the "Ok Google" trigger word during our presentation. However, after further investigation, we found out it was a false positive where the word before "Ok Google" misactivated the device, but the Arduino did not sense the light quick enough to pause the audio of the next word. Despite this negative result, it may still be interesting to conduct a cross-activation measurement study to confirm.

	Trigger Word	Non-Trigger Word
Detected	551	71
Undetected	48	89 046

Table 2: Confusion matrix for the Amazon Echo Dot light activations of trigger word "Hey Alexa" and non-trigger words in the common words experiment. We found a specificity of 0.886 and a sensitivity of 0.920.

	Trigger Word	Non-Trigger Word
Detected	202	1
Undetected	7	29 689

Table 3: Confusion matrix for the Google Home Mini activations of trigger word "Ok Google" in all experiments. We found a specificity of 0.995 and a sensitivity of 0.967.

5.1.2 First Two Volumes of *Clarissa* Dataset

Performing one trial of the experiment, we only found one activation out of 2500 sentences played. The Echo Dot misactivated on the sentence: "no wonder that he has none by any body but himself unexceptionable) has had a hand has the secret pleasure intruded itself,". The server only recorded part of the sentence: "himself unexceptionable". It was flagged as "Audio was not intended for Alexa".

5.1.3 Google Natural Language Questions Dataset

Performing one trial of the experiment, the Echo did not activate on any of the 2500 generated questions.

5.2 Google Home Mini

5.2.1 Top 10k Most Common English Words Dataset

In our experiments, the Google Home Mini effectively did not misactivate due to the words spoken. Every time the trigger word was spoken, the Home Mini activated; for all non-trigger words, the Home Mini would not misactivate. Based on the data records kept by Google's servers, none of the audio for which a request was not responded to was ostensibly not kept, with only a note of the time of occurrence and a generic note ("Used Assistant") with no identifying information or transcript.

5.2.2 First Two Volumes of *Clarissa* Dataset

We did not find any mis-activations of the Home Mini from playing this dataset.

5.2.3 Google Natural Language Questions Dataset

Among the 2500 generated questions played, we found only 1 misactivation. The Home Mini activated while the speaker was playing the segment: "when did the bad place in prehistory because it my way back to scranton?".

In general, we found few misactivations of the Home Mini across these different datasets. We believe there are a few possible reasons for this, compared to the Amazon Echo Dot:

- Firstly, it could simply be that Google has a better training dataset and are better than Amazon at training and evaluating their machine learning models.
- Most of our audio generation pipeline used multiple Google tools: our 10k most common words list was from Google, as was the Natural Questions dataset, and we generated audio from text using Google Translate's text-to-speech engine. It is possible Google has trained their Google Assistant to be more resistant against its own voices for non-trigger words.
- The exact procedure that Google uses is different from that of Amazon. For example, from our observations, we found that if Google decided a clip of audio was not a real trigger word, it would not respond and would not record the audio in its server records, whereas Amazon would still often record audio, even if not intended for Alexa. There may be other places were the sequence of operations that Amazon uses is instrinsically less robust.

5.3 Network Analysis

After examining the network packets sent by the Echo Dot and Google Home Mini at both idle and activated statuses, we found that both devices will communicate with the server every 15 seconds for Echo Dot and every 5 seconds for the Google Home Mini. For the Echo Dot, the network traffic for an activation seems to send four packets with size 1500 bytes and then a packet with reassemble information. Those packages likely contain the audio clips for an activation. The server may or may not give a response depending on the command given to the speaker. For the Google Home Mini, the activation would lead the speaker to first perform a DNS query on www.google.com, then communicate with the given IP address via the QUIC protocol with a lot of packets. Based on that, we can assume that if the length of the data that has been transferred within a few seconds is greater than a certain threshold, then there is an activation. Even if the speaker does not consider it as an activation, the audio clips are sent to the server through internet which potentially exposes sensitive information on the internet. During the analysis, both the Echo Dot and Google Home Mini activation results based on network analysis match the activation identified by light sensing and the server. However, there are a few more "activations" found through network packet analysis. One of the possible reasons for that could be the speaker found a potential activation audio, and it records the audio send to server for further identification. The server denies that the audio clip is an activation so the speaker does not react with turning on the light. However, the audio clip is still sent to the server through internet.

6 Future Work

There are several future paths for this work. Given the time constraints on this project, we were not able to perform as many experiments as we hoped. In the future, we would like to perform experiments with more accents and incorporate more words and phrases. We would also like to test whether different speaker accents have an effect on the number of misactivations when speaking the sentences generated by our 8-gram model. Furthermore, augmenting the audio of the spoken words or phrases is another area that we would like to look into. These augmentations can include varying the pitch of the word/phrase, varying the speed at which the word/phrase is said, injecting noises into the word/phrase, or modifying the frequency spectrum of the audio clip. We can then test whether applying these audio augmentations with different accents would affect the number of misactivations that occur. This also may lend itself to a potentially effective attack: we can try combining an inaudible trigger word and an audible non-trigger word to create subtle misactivations.

7 Conclusion

In this paper, we examine the effects that different accents of speakers have on the number of misactivations for certain IoT devices. We performed experiments with individual words where the speaker spoke the words with an American, British, and Indian accent. Due to time constraints, we were not able to vary the accent of the speaker for the experiments with the sentences and questions. For the Amazon Echo Dot, even though the Amazon server flags a lot of misactivation words as "Audio was not intended for Alexa", there were some misac-

tivation words that were treated as valid trigger words which can be quite dangerous. Having sensitive information said after a misactivated word is treated as a valid trigger word can expose that person's privacy. Moreover, we have not found any patterns in misactivations from the Google Home Mini. Further experiments need to be done for the Google Home Mini and performing experiments with augmenting different words/phrases can help give us more insight on the patterns of words that misactivate certain IoT devices.

References

[1] DUBOIS, D. J., KOLCUN, R., MANDALARI, A. M., PARACHA, M. T., CHOFFNES, D., AND HADDADI, H. When speakers are all ears: Characterizing misactivations of iot smart speakers. *Proceedings on Privacy Enhancing Technologies 2020*, 4 (2020).

[2] GOOGLE. google-10000-english.

[3] KWIATKOWSKI, T., PALOMAKI, J., REDFIELD, O., COLLINS, M., PARIKH, A., ALBERTI, C., EPSTEIN, D., POLOSUKHIN, I., KELCEY, M., DEVLIN, J., LEE, K., TOUTANOVA, K. N., JONES, L., CHANG, M.-W., DAI, A., USZKOREIT, J., LE, Q., AND PETROV, S. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics* (2019).

[4] RICHARDSON, S. Clarissa harlowe; or the history of a young lady. Project Gutenberg. [Online].

[5] SCHÖNHERR, L., GOLLA, M., EISENHOFER, T., WIELE, J., KOLOSSA, D., AND HOLZ, T. Unacceptable, where is my privacy? exploring accidental triggers of smart speakers. *arXiv preprint arXiv:2008.00508* (2020).

[6] SUGAWARA, T., CYR, B., RAMPAZZI, S., GENKIN, D., AND FU, K. Light commands: laser-based audio injection attacks on voice-controllable systems. In *29th {USENIX} Security Symposium ({USENIX} Security 20)* (2020), pp. 2631–2648.

[7] ZARANDY, A., SHUMAILOV, I., AND ANDERSON, R. Hey alexa what did i just type? decoding smartphone sounds with a voice assistant. *arXiv preprint arXiv:2012.00687* (2020).

Notes

Our Github repository containing both the code and data can be found here: https://github.com/wongmaster3/Speaker-Misactivations.