Daniel J. Dubois*, Roman Kolcun, Anna Maria Mandalari, Muhammad Talha Paracha, David Choffnes, and Hamed Haddadi

When Speakers Are All Ears: Characterizing Misactivations of IoT Smart Speakers

Abstract:
Internet-connected voice-controlled speakers, also known as *smart speakers*, are increasingly popular due to their convenience for everyday tasks such as asking about the weather forecast or playing music. However, such convenience comes with privacy risks: smart speakers need to constantly listen in order to activate when the "wake word" is spoken, and are known to transmit audio from their environment and record it on cloud servers. In particular, this paper focuses on the privacy risk from smart speaker *misactivations*, *i.e.*, when they activate, transmit, and/or record audio from their environment when the wake word is *not* spoken. To enable repeatable, scalable experiments for exposing smart speakers to conversations that do not contain wake words, we turn to playing audio from popular TV shows from diverse genres. After playing two rounds of 134 hours of content from 12 TV shows near popular smart speakers in both the US and in the UK, we observed cases of 0.95 misactivations per hour, or 1.43 times for every 10,000 words spoken, with some devices having 10% of their misactivation durations lasting at least 10 seconds. We characterize the sources of such misactivations and their implications for consumers, and discuss potential mitigations.

Keywords: smart speakers, voice assistants, privacy, IoT, voice command, voice recording, wake word

DOI Editor to enter DOI
Received ..; revised ..; accepted ...

***Corresponding Author: Daniel J. Dubois:** Northeastern University, E-mail: d.dubois@northeastern.edu
Roman Kolcun: Imperial College London, E-mail: roman.kolcun@imperial.ac.uk
Anna Maria Mandalari: Imperial College London, E-mail: anna-maria.mandalari@imperial.ac.uk
Muhammad Talha Paracha: Northeastern University, E-mail: paracha.m@husky.neu.edu
David Choffnes: Northeastern University, E-mail: choffnes@ccs.neu.edu
Hamed Haddadi: Imperial College London, E-mail: h.haddadi@imperial.ac.uk

1 Introduction

Internet-connected voice-controlled speakers, also known as *smart speakers*, are popular IoT devices that give their users access to voice assistants such as Amazon's Alexa [1], Google Assistant [2], Apple's Siri [3], and Microsoft's Cortana [4]. Smart speakers are becoming increasingly pervasive in homes, offices, and public spaces, in part due to the convenience of issuing voice commands [5]. This allows users to perform Internet searches, control home automation, play media content, shop online, *etc.*—by saying a "wake word" (*e.g.*, "*Alexa*") followed by a question or command.

However, this convenience comes with privacy risks: smart speakers need to constantly listen in order to activate when the "wake word" is spoken, and are known to transmit audio from their environment and record it on cloud servers. In particular, in this paper we focus on the privacy risk from smart speaker *misactivations*, *i.e.*, when they activate, transmit, and/or record audio from their environment when the wake word is *not* spoken.

Several reports demonstrated that such misactivations occur, potentially record sensitive data, and this data has been exposed to third parties. For instance, the Google Home Mini bug that led to transmitting audio to Google servers continuously [6], multiple voice assistant platforms outsourced transcription of recordings to contractors and some of these recordings contained private and intimate interactions [7]. There have been other anecdotal reports of everyday words in normal conversation being mistaken for wake words [8].

In this work, we conduct the first repeatable, controlled experiments to shed light on what causes smart speakers to misactivate and record audio. In particular, we detect when smart speakers misactivate, which audio triggers such misactivations, how long they last, and how this varies depending on the source of audio, time, and the country where the smart speaker is deployed.

To achieve these goals, a key challenge is to determine how to expose smart speakers to representative spoken dialogue in a repeatable, scalable way. While this could potentially be accomplished using researchers who speak from scripts, this is not scalable and represents a

BY-NC-ND

Fig. 1. Testbed: camera and light to capture activations (top left), speakers controlled by us to play audio content from TV shows (center left), smart speakers under test (right). The bottom left shows the frame of an activation, *i.e.*, an Echo Dot device lighting up, as detected by the camera.

small number of voices. Instead, we rely on a key insight: popular TV shows contain reasonably large amounts of dialogue from a diverse set of actors, and spoken audio can be tied to a text transcript via closed captions. In this paper, our experiments use Netflix content for a variety of themes/genres, and we repeat the tests multiple times to understand which non-wake words consistently lead to unwanted activations (*i.e.*, misactivations).

Another key challenge is determining when a misactivation occurs. To address this, we use a multi-pronged approach composed of capturing video feeds of the devices (to detect lighting up when activated), capturing and analyzing network traffic (to detect audio data sent to the cloud), and data about recordings provided by smart speakers' cloud services (when available).

To conduct this research, we built a custom testbed (see Fig.1) and used it to expose seven smart speakers, powered by four popular voice assistants, to 134 hours (more than a million words) of audio content from 12 popular TV shows. We conducted these same experiments in the US and UK, and repeated the experiments two times in each location, for a total of 536 hours of playing time. We found that most misactivations are not repeatable across our initial two rounds of experiments, *i.e.*, some audio does not consistently trigger a misactivation. We conduct additional experiments for the misactivations in the first two rounds, playing the audio 10 more times. Our key findings are as follows:

- Smart speakers exhibited up to 0.95 misactivations per hour, or 1.43 every 10,000 words spoken. While the good news is that they are not constantly misactivating, misactivations occur rather frequently.
- The duration of most misactivations is short, but, for some devices, 10% of the misactivations are 10 seconds long or more. The latter is enough time to record sensitive parts of private conversations.
- Misactivations are often non-deterministic, *i.e.*, a smart speaker does not *always* misactivate when exposed to the same input.
- We found some of the text leading to misactivations to be similar in sound to the wake words. We did not find any evidence of undocumented wake words. For many misactivations, there was no clear similarity to wake words, meaning there is a high potential for recording audio unrelated to voice commands.
- When comparing misactivations from the US and the UK, there are often more misactivations in the US than the UK, and the overlap is small. This may indicate that smart speakers in different countries use different voice recognition models, or that they are sensitive to the acoustic properties of different test environments.

To summarize, our key contributions are as follows: (*i*) we conduct the first repeatable, at-scale study of misactivations from smart speakers, (*ii*) we used a new combination of video processing, network traffic analysis, and audio content analysis to robustly identify misactivations and the dialogue that triggered them, and (*iii*) we analyzed the data from these experiments to characterize the nature of such activations and their implications. To foster research in the relatively new field of smart speaker analysis, we make all of our testbed code and data available at `https://moniotrlab.ccis.neu.edu/smart-speakers-study`.

2 Assumptions and Goals

In this section we state the assumptions, threat model, and goals of this work.

2.1 Assumptions and Definitions

Smart speakers and voice assistants. We define smart speakers as Internet-connected devices equipped with a speaker, a microphone, and an integration with a cloud-enabled service (*i.e.*, voice assistant) responsible for receiving and interpreting voice commands when they are activated (see Fig. 2). We assume that smart speakers are constantly processing the data from their microphone (locally), which they use to decide whether to trigger an activation or not. Each smart speaker is

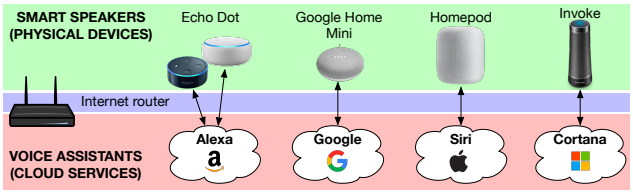


Fig. 2. Relationship between smart speakers and voice assistants.

associated with one or more *wake words*, which are the only words that are intended to trigger an activation.

Activations and Misactivations. We define activation every time a smart speaker perceives a wake word and listens for a subsequent command. This may or may not be followed by sending data from its microphone over the Internet to its cloud-enabled voice assistant. If an activation is not triggered by the wake word, such activation is an unauthorized activation, and defined as *misactivation*. The duration of an activation is defined as the period between we detect an activation signal and when such signal is no longer detected.

Activation signals. We define activation signal any signal that a smart speaker emits when it is activated. These signals can be used as a way to detect if a smart speaker has been activated and for how long. The signals we consider for this work are: (*i*) *visual activation*, available to all the smart speakers we consider, which is the result of some LEDs lighting up on top of the device with a particular activation pattern; (*ii*) *cloud activation*, available only for Amazon and Google smart speakers, which is the result of adding entries to the log of activations stored on the cloud and user accessible; (*iii*) *network traffic activation*, available to all smart speakers (since all of them need to communicate over the Internet), which is the result of the actual transfer of a voice recording over the Internet.

Our activation signals provide evidence of three smart speaker activities: interpreting audio locally, transmitting audio over the Internet, and recording audio on remote servers. A visual activation signal indicates that the device is interpreting audio, a network traffic activation indicates that a device is transmitting a recording of audio from its environment, and the cloud activation signal indicates that the audio has been recorded by the voice assistant service.

2.2 Threat Model

We consider the following threat model for our analysis.

Victim. The victims are any persons in microphone range of smart speakers, who expect that the only audio transmitted over the Internet corresponds to voice commands preceded by a wake word.

Adversary. The adversary is any party that can access audio recorded from smart speakers. Examples include the voice assistant provider and its employees, any other company or contractor that the company shares recordings with, and any party that can gain unauthorized access to these recordings.

Threat. The primary threat we consider is audio data (*e.g.*, conversations or other sensitive information) recorded due to misactivations, thus exposing recordings without authorization from the victims to the adversary.

2.3 Goals and Research Questions

The main goal of this work is to analyze the potential privacy risks arising from smart speakers misactivations. In particular, this work answers the following research questions (RQ):

1. *How frequently do smart speakers misactivate?*
   We characterize how often and consistently a smart speaker misactivates when exposed to conversations. The more misactivations occur, the higher the risk of unexpected audio recordings.
2. *How well do our three activation signals correlate?*
   Smart speakers light up when activated (to inform users they are actively listening for a command), and in many cases they are expected to transmit a recording of audio to their cloud servers for analysis and storage. We study whether we see consistency in the activation signals and whether, for example, there is evidence of audio transmission that is not reported as recorded by the cloud service.
3. *Do smart speakers adapt to observed audio and change whether they activate in response to certain audio over time?*
   We track whether smart speakers are more or less likely to misactivate when exposed to the same (non-wake-word) audio over time. More misactivations indicate higher risk of unexpected recording, while fewer misactivations reduce this risk. Changes in misactivation rates can occur if voice assistant providers update their machine-learning models for interpreting wake words over time.
4. *Are misactivations long enough to record sensitive audio from the environment?*
   A long misactivation (*i.e.*, a possible long voice recording) poses a higher privacy risk than a short one since it provides more data (*e.g.*, context and details of a conversation) to the adversary.

5. *Are there specific TV shows that cause more overall misactivations than others? If so, why?*
   Each TV show we selected has different dialogue characteristics (*e.g.*, word density, accent, context, *etc.*). We measure which shows trigger more misactivations to understand which characteristics correspond to an increased risk of misactivation.
6. *Is there any difference in how smart speakers misactivate between the US and the UK?*
   All the smart speakers we consider (except for Invoke) are available in both the US and the UK. We study whether the region in which the smart speakers are sold and deployed has any effect on the risk of triggering misactivations.
7. *What kind of non-wake words consistently cause misactivations?*
   Consistent misactivations in response to certain words or audio is useful for understanding areas for improvement in smart speakers, words to avoid for privacy-conscious consumers, and any potential undocumented wake words/sounds.

In the next section, we present details of the methods we use for gathering data to answer these questions. We then analyze the data to answer these questions in §4 and provide a discussion of the results in §5.

3 Methodology

This section describes the methods we use to detect and characterize smart speaker misactivations (summarized in Fig. 3). We start by selecting a set of popular smart speakers (§3.1) and a diverse set of television-show audio (§3.2) to play at them. We then use our testbeds (§3.3) to perform the experiments by playing the television shows' audio tracks (§3.4). We then conduct analysis to recognize activations, and distinguish legitimate ones from misactivations (§3.5). In §3.6, we describe how we analyze each misactivations to determine their cause and privacy implications. The remainder of this section will explain the details of our approach.

3.1 Smart Speakers and Voice Assistants

We selected smart speakers based on their popularity, global availability, and the different voice assistants that power them. Specifically, we tested the following:

TV Show	Time	# of words	Words/Time
Dear White People	14h	100K	119 wpm
Friday Night Tykes	11h	110K	167 wpm
Gilmore Girls	11h	117K	177 wpm
Greenleaf	10h	64K	107 wpm
Grey's Anatomy	11h	99K	150 wpm
Jane The Virgin	11h	87K	132 wpm
Narcos	11h	50K	76 wpm
Riverdale	11h	70K	106 wpm
The Big Bang Theory	10h	83K	138 wpm
The L Word	11h	73K	111 wpm
The Office (U.S.)	12h	108K	150 wpm
The West Wing	11h	96K	145 wpm
SUM	134h	1057K	131 wpm

Table 1. Content used for playing audio. List of TV shows, the playing time, the amount of dialogue in terms of words, and the density of dialogue in words per minute (wpm).

- *Google Home Mini* powered by Google Assistant (wake word: "OK/Hey Google");
- *Apple Homepod* powered by Apple's Siri (wake word: "Hey Siri");
- *Amazon Echo Dot* 2^{nd} *and* 3^{rd} *generation* powered by Amazon's Alexa (alternative wake words: "Alexa", "Amazon", "Echo", "Computer").
- *Harman Kardon Invoke* powered by Microsoft's Cortana (wake word: "Cortana") (US only);

We use two generations of Echo Dot devices because of the different microphone configuration (seven microphones in the case of the 2^{nd} gen. compared to four in the 3^{rd} gen.), which may affect their voice recognition accuracy. We tested multiple devices at the same time, all placed next to each other, and set the volume to the minimum level to minimize interference. Since all the devices we tested use different wake words, we expected simultaneous activations of multiple devices to be rare. Indeed, we did not observe any case where multiple smart speakers activated simultaneously, and avoided any cases where one smart speaker's response activates another. Due to the variety of wake words supported by Amazon devices, we use two devices for each of the 2^{nd} and 3^{rd} generations (four in total) so that we can test four wake words for each round of experiments. The Harman Kardon Invoke was present in the US testbed only, due to the fact that Harman Kardon has not introduced this device to the UK market.

3.2 Audio Content

A key challenge for characterizing smart speaker activations is identifying a way to expose these devices to representative spoken dialogue in a repeatable, scalable

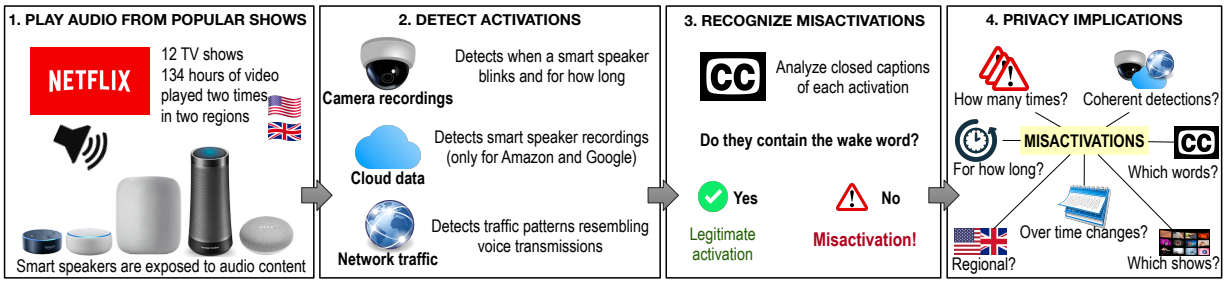


Fig. 3. Overview of the approach: our pipeline for detecting and characterizing misactivations from the smart speaker testbed.

way. While this could potentially be accomplished using researchers who speak from scripts, this is not scalable in terms of cost and time with respect to the number of voices and diversity of people. Another approach we investigated is using audiobooks; however, in general written text does not use the same language as spoken dialogue, and there is usually one narrator speaking. Instead, we rely on a key insight: popular TV shows contain large amounts of dialogue from a diverse set of actors, often more than 10 per show, and spoken audio can be tied to a text transcript via closed captions for automated analysis. While there are many sources of TV shows, we used content from the Netflix streaming platform. One advantage of Netflix is the wide range of contents and the availability of subtitles for each of them. Note that we ensured that all experiments did not contain any interruptions to the audio stream.

We selected 12 shows that include popular and diverse features such as density of dialogue, cultural context, genre, technical language, *etc.* (Tab. 1). While the total duration of episodes varies according to the show, we played more than 10 hours of content from each show, starting from the most recent available episode. In total, we tested 217 episodes, totaling 134 hours and 1.057 million words of content.

3.3 Testbed Description

Our testbed is composed of the following components: (*i*) the smart speakers under test (described above); (*ii*) a pair of additional non-smart speakers used to play audio from the TV shows; (*iii*) an RTSP-compatible camera for recording a video and audio feed from all the speakers; (*iv*) a TP-Link smart plug, connected to the smart speakers, which is able to programmatically cycle their power; (*v*) a *coordinating server*.

The coordinating server manages the lifecycle of each experiment and it is responsible for acting as router and providing Wi-Fi connectivity to the smart speakers so that they are isolated from other networks and devices (to avoid interference). It also plays the audio tracks from TV shows through the pair of non-smart speakers, it captures the network traffic using tcpdump, and finally it records video from the camera using RTSP.

All the smart speakers are placed in a cabinet that is insulated from external sounds, with the camera on top, and the non-smart speakers next to them (see Fig. 1). The volume for playing the video content through the non-smart speakers is set to such a level that it, in our opinion, resembles the volume of a casual conversation. The output volume of the smart speakers was set to the minimum level to reduce the impact of possible voice responses from a (mis)activated speaker.

We built two instances of this testbed and deployed one in the US and one in the UK. This provides data to inform whether there are any regional differences in smart speaker behavior in response to identical stimuli.

3.4 Experiments

Preparation. To prepare the smart speakers for the experiments, all of them are reset to factory settings and associated with newly created accounts specifically used for this purpose only. This is important as some devices might try to create a profile of the voices they hear, and thus skew the results of our experiments (we find evidence of this in §4.3). Our experiments represent "out of the box" behavior for smart speakers over the course of a few months, but do not represent behavior over long periods of time (*e.g.*, many months or years).
Lifecycle of each experiment. Each experiment consisted of playing all the 134 hours content, each episode at the time. We first turn off all the smart speakers, wait 5 s, turn them on, and then wait 90 s to ensure they have booted and are ready to be used. Then, we start the following all at the same time: (*i*) play audio from

an episode, (*ii*) start camera recording using FFMPEG, (*iii*) start network traffic recording on the router using an instance of tcpdump for each smart speaker. When the played episode is over, we stop the recording of the traffic and the camera, and a new cycle can start.

Repetition of experiments. We repeat all the experiments two times for each device and wake word, reaching a total of 268 hours of played content to all the devices. Moreover, we repeat all the two experiments in two testbeds (in the US and in the UK), totaling 536 hours of cumulative playing time.

Additional consistency experiments. Misactivations may occur nondeterministically, and may change as voice assistant platforms update their logic/models for detecting wake words. To test for this, we perform 10 additional experiments to identify how *consistently* misactivations occur. The difference between base experiments and consistency experiments is that in base experiment we play all the television shows, while in the consistency experiments we only play the portion of the content that triggered a misactivation on a base experiment (plus 20 s before and after, to ensure to replay all the content that triggered such misactivation). For this reason consistency experiments take a fraction of the time compared to base ones. In the case of Invoke, we could only perform two rounds of confirmatory experiments instead of ten because of a service outage not resolved in time for this study.

Special consideration for Echo Dot devices. Recall that users can select from four wake words on Amazon devices. Given that we have only two of each generation of Echo Dot devices, and that we treat each generation as separate types of devices, we need to run additional tests to cover all wake words. Namely, we run one set of two experiments with two different wake words on each device, then conduct an additional set of two experiments using the other two wake words.

3.5 Detection of (Mis)activations

We detect activations using the camera (based on a device lighting up), cloud (based on the information about recordings accessible via app or web page from the voice assistant provider, currently available only for the Alexa and Google voice assistants), and network traffic (based on traffic peaks). For any detected activation, we search for a wake word in the closed captions of the part of the television show that caused the activation. If the wake word was spoken, the activation is ignored (*i.e.*, not considered in this analysis because it is an expected activation), otherwise it is labeled as a misactivation.

Note that for an activation to be considered a misactivation, its subtitles at the time of the activation (and 20 s before/after, to provide tolerance for lag between subtitle appearance and words being spoken) must contain the wake word with its exact spelling, but ignoring the capitalization. For example, “alexa” and “HEY SIRI” are valid wake words for Alexa and Siri, while “Alexandria” and “Siri” (without “hey”) are not.

3.5.1 Detection from Camera

While the audio content is played, a fixed camera records a video of each smart speaker under test. The camera is positioned such that all smart speakers are visible in the video stream. For all the devices we tested, a light at the top of the smart speaker turns on to indicate an activation (see the bottom left part of Fig. 1). We process the video stream from the camera to find instances of these lights turning on. While image recognition in general is a hard problem, identifying activation lights is much more straightforward. Namely, we compare each frame of video with a reference image (containing all the devices in non-activated state), and differences indicate the light turning on. Further, the smart speakers remain in a fixed position throughout the experiment, so the coordinates of the pixels where the change in color is detected reveal which devices activate. We additionally use this detection method to measure the duration of an activation, which we define as the time between when the light turns on and when it subsequently turns off. Finally, we manually reviewed all camera-detected activations in our study and removed from the data anomalous situations in which the device is not lighting up or is signaling a problem (such as blinking red). As such, all camera-detected activations were verified to be true activations.

3.5.2 Detection from Cloud

According to manufacturer documentation [9, 10], every time a device powered by Amazon Alexa or by Google Assistant is activated, the activation is recorded on the respective cloud servers. Both Amazon and Google provide a web interface for accessing the list of activations, as well as the date and time at which they occurred. We assume that all cloud-reported activations represent real activations (legitimate or misactivations), and thus can

Device	Destination (D)	B_{max} [KB/s]	A_{min} [KB/s]	Threshold (X) [KB/s]
Google Home Mini	*google.com	41.69	50.483	46.086
Homepod	*apple.com	145.855	6.878	76.366
Echo Dot (2^{nd} & 3^{rd} gen.)	bob-dispatch-* *.amazon.com	8.889	15.455	12.172
Invoke	*msedge.net, *skype.com *microsoft.com, *bing.com	9.728	17.408	13.568

Table 2. Network traffic detection thresholds per device: Each threshold represents the peak 1-second outbound throughput sample during a 20-second window.

use this as a form of ground truth for measuring the accuracy of the camera detection. Namely, a cloud activation is proof that a recording actually occurred, that it has been sent over the Internet, and stored on a voice assistant provider's server. Note that while we assume all cloud activations are real activations, we do not assume that all activations are reported in the cloud.

3.5.3 Detection from Traffic

To identify whether audio recordings from smart speakers are transmitted over the Internet, we analyze network traffic from each device. Note that all activation traffic was encrypted and there was no destination that received only audio transmissions, meaning we cannot simply look for keywords in network traffic or domain names. Thus, the key challenge for using network traffic as a signal for audio transmission is distinguishing this transmission from other "background" network traffic unrelated to activations.

To address this challenge, we conduct a set of controlled experiments consisting of idle periods and activation periods. For the idle periods (when the smart speaker is not in use), we capture network traffic and label it as *background traffic*. Next, we observe network traffic changes when the device is activated using the wake word, and label this as *activation traffic*.

To detect background traffic, we collected the network traffic from every smart speaker for a week, without playing any content. To obtain samples of activation traffic, we played a short simple question such as "What is the capital of Italy?" to every device. The duration of each *activation experiment* was $W = 20s$, which is sufficient time to ask a question and wait for the reply. These activation experiments were repeated at least 135 times for each smart speaker.

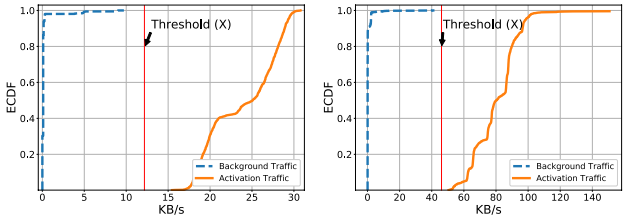
By comparing the background traffic with the activation traffic we can identify unique characteristics of the network traffic that can be used to detect the activation of a device. In particular, we inferred the destination domain(s), protocol, and port that are contacted during activations, and a threshold X, which represents the amount of traffic sent to such destination (in bytes/second). The thresholds and the other parameters of our activation detection heuristics are reported in Tab. 2, and were identified as follows.

Measuring the activation threshold X. To measure this threshold, we perform these steps for each device:

1. Empirically determine the destination D of the activation traffic by looking at samples. D includes the domain name, protocol, port, and must be contacted in every activation experiment.
2. We calculate B_{max}, defined as the maximum amount of data (in bytes/second) sent to D over every possible time window of length W in the background traffic. This serves as a "high pass filter" since we are interested in peaks of network traffic caused by audio transmission.
3. Given S as the set of samples of activation traffic, we measure A_i, $i \in S$, which is the maximum amount of data (in bytes/second) sent to D during each activation traffic sample (that is always of length W). We refer to A_{min} as the minimum value among A_i for all the samples.
4. We calculate the threshold $X = \frac{A_{min}+B_{max}}{2}$, which is the average between the largest peak of traffic volume to D in background traffic in a time window W, and the minimum peak traffic volume sent to D during an activation. The idea behind this definition is to provide a balance between false negatives and false positives for detecting an activation.

Detecting activations. To detect an activation, we use the destination D and the threshold X in the following way. Given unlabeled traffic as input, we calculate U_t, where $t = 1 \dots T$ is a time index and T is the duration of the pcap file (in seconds). U_t is the amount of traffic (in bytes/second) sent to D. We then find every case in which $U_t > X$: such cases are activations happening at time t. For the sake of this study, we consider activations happening within 5 s as the same activation.

Fig. 4 depicts, for the Echo Dot 3^{rd}generation and Google Home Mini, an ECDF of A_i from all the activation experiments and B_{max} value for every possible window of size W in the background traffic. In this case $A_{min} > B_{max}$ and thus the heuristics for detecting an activation is expected to have few false positives and negatives because we never see background traffic peaks that are larger than activation traffic peaks. We show statistics for other devices in Tab. 2.



(a) Echo Dot 3^{rd} generation. **(b)** Google Home Mini.

Fig. 4. ECDF for background and activation traffic: the plot shows a clear distinction between the amount of traffic sent when an activation is triggered and when the device is not in use.

We assume that audio transmission from smart speakers results in peaks of throughput beyond those from idle periods. When this is not true, such in the case for Homepod (case of $A_{min} < B_{max}$ in Tab. 2), our approach is more likely to either miss activations, or falsely detect activations. For instance, transmissions of short recordings may not generate enough traffic volume to be detectable beyond background traffic. Conversely, background traffic may exhibit peaks that are misinterpreted as activations. Future updates to devices may also change their background and activation traffic, so our thresholds may need to be revisited in the future.

3.6 Determining Privacy Implications

After identifying misactivations, we analyze the data to understand their privacy implications. To this end, we measure several characteristics of each misactivation:

- whether it appears in all the experiments or not, to determine if privacy exposure happens consistently or nondeterministically;
- if the duration of an activation is long enough to violate the privacy of a conversation;
- whether rate of misactivations change over time, potentially indicating that the voice assistant provider is creating a voice profile of its users, thus collecting private information about them;
- whether the different detection methods yield similar results, in part to verify if the smart speaker is correctly signaling to its user that it is recording;
- whether different types of voice content have a higher (or lower) probability of causing a misactivation and therefore exposes certain voices or types of conversations to different privacy risks;
- determine if geographic location (US or UK) changes the misactivation rate and duration, and thus the amount of privacy exposure;
- determine which words cause the most misactivations, to understand the existence of any undocumented "wake words" aimed at capturing a particular type of conversations or sounds.

3.7 Limitations

Deployment environment. Our experiments were conducted in a cabinet to isolate the speakers from outside noise in a shared space, which causes differences in results compared to more common deployment scenarios such as a tabletop in an open room, or on a bookshelf. To understand the impact of this limitation, we ran a small-scale experiment with the smart speakers outside the cabinet and did not observe any significant difference in activations. In addition, the cabinets and room layout are different between the US and the UK testbed, adding environmental differences when comparing US and UK misactivations.

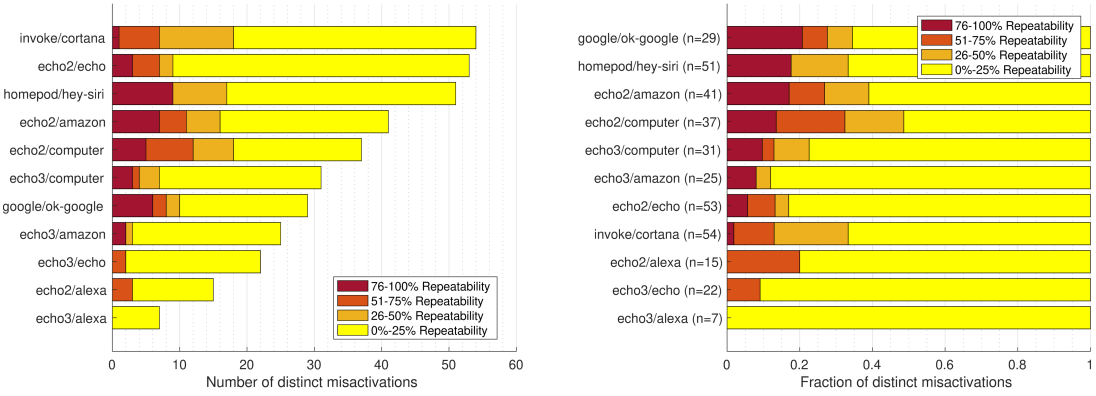
Single user *vs.* multiple users. We exposed the smart speakers to voice material that includes hundreds of voices, so that our experiments could represent an extreme multi-user scenario. For this reason, our analysis cannot distinguish the implications between a smart speaker always used by a single user *vs.* multiple users.

4 Misactivations Analysis

In this section, we analyze the misactivations detected according to the methodology we explained in §3 to answer our research questions (see §2.3). We run our two base experiments in Nov. 2019 (US) and Jan. 2020 (UK), while the 10 consistency experiments were run in the US only on Feb. 2020 (except for Invoke, which has two confirmatory experiments run in Dec. 2019).

Throughout the section, we consider the union of activations across the two base experiments, so that activations appearing in multiple experiments from the same device at the same time (with a tolerance of 5 s) are counted only once. We refer to such misactivations as *distinct misactivations*.

Unless otherwise specified, we consider only misactivations detected by camera. We could not use the other detection methods in all analyses because for Invoke and Homepod, cloud detection is not available and traffic detection has significant false positives. For camera activations, we avoided any false positives by manually checking *all* of them.



(a) Number of misactivations by device.

(b) Repeatability of misactivations by device.

Fig. 5. Misactivations overview: number of overall misactivations and their repeatability across our multiple experiments.

We were unable to verify the absence of false negatives, and thus this study provides a lower bound on privacy exposure from misactivations. To better understand and support the privacy implications of focusing on camera-detected activations, in §4.2 we compare the results across all the different detection methods.

Finally, we discuss the differences between the US and UK testbed in a dedicated section §4.6; all other parts of this section focus on misactivations detected only from the US testbed.

4.1 Misactivations and their Repeatability

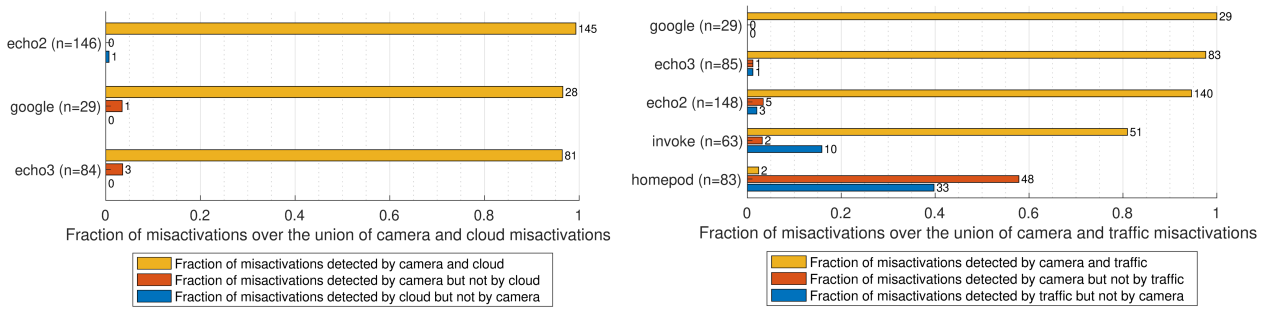
We now answer the question of how many times the tested smart speakers misactivate (RQ1). We assume that a higher rate of misactivations means a higher rate of recording audio from the environment, thus exposing the privacy of its user more frequently. Fig. 5a shows the number of distinct misactivations for all the speakers, including—for Amazon devices—any combination of generation (*i.e.*, Echo Dot 2^{nd} gen. and 3^{rd} gen.) and wake word (*i.e.*, Alexa, Amazon, Computer, Echo).

The figure shows that Invoke misactivates the most times (54), followed by Echo Dot 2^{nd} gen. (wake word "Echo") and Homepod (53 and 51 misactivations). However, we observed that the majority of misactivations do not appear in every round of our experiments. We measured this by calculating the *repeatability* of a misactivation, which is defined as the number of experiments in which the misactivation is detected, divided by the total number of experiments (*i.e.*, base experiments plus consistency experiments). Fig. 5a shows the quartiles distribution of misactivations in absolute numbers, while Fig. 5b shows it in relative terms. By looking at both figures we can see that for all the devices, the majority of misactivations have a low repeatability, meaning that the most common case is that they are detected in less than 25% of the experiments; however, we notice also a significant amount of cases where misactivations are consistent for a large majority of the experiments (75% or more). This is most evident for Google Home Mini, where 20.7% misactivations have $> 75\%$ repeatability, followed by Homepod (with 17.7% highly repeatable misactivations), and then by two Echo Dot 2^{nd} gen. (Amazon and Computer wake words), with respectively 17.1% and 13.5% highly repeatable misactivations.

Takeaways. Devices with the most misactivations (such as Invoke and Echo Dot 2^{nd} gen. "Echo") expose users to unintentional audio recordings more often than devices with less misactivations (such as Echo Dot 2^{nd} and 3^{rd} generation configured with the Alexa wake word). The prevalence of misactivations with low repeatability across all devices suggests that their wake word recognition capability is nondeterministic in many cases of misactivations, since it does not produce consistent results across experiments. From a privacy perspective, this makes it hard to understand why a device misactivates and predict when a device is going to misactivate or not. But this is also a source of concern, since having a device misactivating unpredictably still results in a risk of exposing private audio from the environment to the voice assistant service provider.

4.2 Misactivations by Detection Method

For this analysis our goal is to answer RQ2, *i.e.*, understand if there are any disagreements between misactivation detection via the three methods we consider.



(a) Camera vs. cloud.

(b) Camera vs. traffic.

Fig. 6. Comparison of detection methods: analysis of the differences in the detection capabilities of different detection methods.

Measuring these disagreements first helps us to understand the validity of a detection method. When two different methods agree, this can increase our confidence in the reliability of the detection method (*i.e.*, stronger evidence that such misactivations actually occurred). When two different methods disagree, it could be due to poor detection accuracy (*e.g.*, our camera detection method does not report the device lighting up when it is actually lighting up or vice-versa), or when the smart speaker sends wrong activation signals (*i.e.*, the device lights up when not being activated or vice-versa). Since we do not have ground truth for all activations (which would require access to smart speaker internal state, decrypted network traffic, and cloud servers), we rely on using information from multiple detection methods to decide which one we can consider most reliable.

We begin by comparing our primary detection method (camera) with the cloud. Fig. 6a shows that for devices offering a cloud detection API, the detection of misactivations from both camera and cloud are almost perfectly correlated, with the percent of misactivations detected by both between 96.4% and 99.3%. Note that we cannot conduct this analysis for Invoke and Homepod since they do not offer cloud activation detection.

We then repeated the same analysis using activations detected from network traffic. Fig. 6b shows perfect correlation for Google Home Mini (percent of misactivations detected by both camera and traffic of 100%), and high levels of consistency between Echo Dot 3^{rd} and 2^{nd} gen. (percent of misactivations detected by both camera and traffic of 97.6% and 94.6%). Invoke also shows evidence of consistency, although not as strong, with a percent of misactivations detected by both camera and traffic of 81.0%. We analyzed the consistency between cloud and traffic activations, and found results nearly identical to the ones in Fig. 6b due to the high correlation between camera and cloud.

The reason for the slight disagreements between activation signals may be caused by false negatives or false positives on the traffic detection method, or by the device not correctly reporting its activations by lighting up or by storing them in the cloud. We simply do not have the ground truth to determine the root cause. Interestingly, we found four misactivations detected by the camera and not reported in the cloud recordings. For three of these, we found evidence of audio transmission in network traffic. This may indicate omissions in cloud data (if the transmission was recorded) or that some transmissions are not recorded. Either case has important implications for transparency of device activations.

The misactivations detected from network traffic for the Homepod device show almost no overlap with the camera. This is not surprising given that there are background traffic peak volumes that are larger than those observed during activations (see Tab. 2).

Takeaways. Camera and cloud detection (when available) are almost perfectly correlated, although there are a handful of activations where camera and cloud disagree. We manually inspected the camera feed, and did not find anything that explains this behavior; however, it happens so rarely (four cases of camera activations not detected by cloud and one case of cloud activation not detected by camera), that we do not consider the risk of stealthy audio recording significant. Conversely, we have seen cases of more significant disagreement in misactivations detected from network traffic (*e.g.*, Echo Dot 2^{nd} gen.). This motivates further research—likely using classifiers and additional detection features—to more accurately identify transmission of audio in encrypted network traffic.

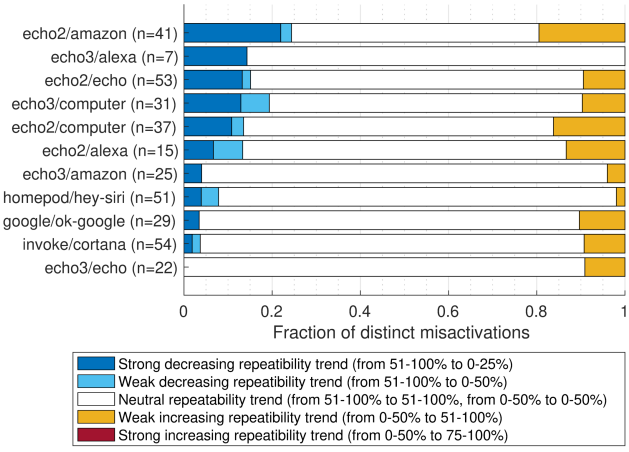


Fig. 7. Misactivation trend: measuring how well past misactivations from our initial batch of two experiments (Nov 2019) are still detected in our confirmatory experiments (Feb 2020 for all devices, except for Invoke: Dec 2019).

4.3 Repeatability Trend

In this analysis we want to answer RQ3, that is understanding if any smart speaker ecosystems adapt their activation behavior over time. This may occur, for example, if an ecosystem builds voice profiles of consumers so that they reduce misactivations over time (*e.g.*, by filtering out other voices such as those from television shows). To answer this question, we determine if there are any trends in the activation repeatability between our main runs of experiments and the 10 (two for Invoke) additional confirmatory experiments. For this study we define trend as follows:

- *Decreasing*: if the repeatability changes from $> 50\%$ in the main experiments to less than $< 50\%$ in the confirmatory experiments. If it becomes even $< 25\%$, we define the trend as *strong decreasing*.
- *Increasing*: if the repeatability changes from $< 50\%$ to $> 50\%$. If it becomes even more $> 75\%$, we define the trend as *strong increasing*. However, we did not find any such case for any devices.
- *Neutral*: if it is neither decreasing nor decreasing.

Fig. 7 shows the repeatability trend for the devices we tested. The figure shows that the vast majority of activations from all speakers have a neutral trend, but also indicates the presence of a strong decreasing repeatability trend for all Echo Dot devices compared to the others. This suggests that Amazon may be adapting its wake word detection approach over time, possibly by building profiles based on user voices and/or the content of their speech. We also observe some degree of weaker trends of both decreasing and increasing repeatability, but we believe it may just be a result of the misactivation nondeterminism (most activations are not repeatable, as shown in §4.1) and the fact that the number of main experiments for all smart speakers (except Invoke) is lower than the confirmatory experiments.

We do not know if the trends we see are caused by the presence of multiple users, by the content of their dialogue, a combination of both, or some other factor. Answering this would require a combination of (proprietary) information about smart speaker implementations and a large suite of additional tests.

Takeaways. Our results provide evidence that Echo Dot devices are adapting to observed audio, based on the relatively prominent strong decreasing trend in activations when compared to other devices. It is, of course, possible that other voice assistant ecosystems that we tested also adapt over time; however, we did not see strong evidence of it during our tests.

An important question is how Echo Dot devices are able to filter out misactivations after repeatedly hearing the same audio. We hypothesize that they may retrain their voice recognition model based on the voice characteristics and the content of speech. However, by doing so there is a potential to also build a voice profile that can be used to identify an individual user. While such a voice profile can itself constitute a privacy risk, it can also be used to mitigate privacy risks in terms of reducing the number of misactivations leading to recordings.

4.4 Duration of Misactivations

We now answer RQ4, and thus understand what is the duration for each activation. A longer duration means a greater exposure of privacy since more conversation is likely to be transmitted. Since all the devices we have signal their activation activity (and intention to record) by lighting up, we assume that the duration of a misactivation is the amount of time a device is lit up.

Fig. 8 shows how the duration of misactivations is distributed in terms of percentiles. In the median case, we see a misactivation duration of up to 4 s in the case of Homepod and Echo Dot 2^{nd} gen (Alexa and Computer wake words). In the less common 75^{th} percentile case (P75), we observe the largest misactivation duration of 7 s for the Homepod, followed by Echo Dot 2^{nd} gen. with 6 s (Computer and Amazon wake words). Finally,

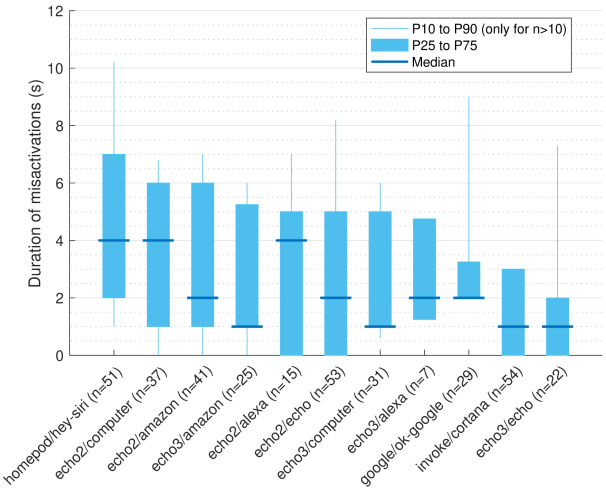


Fig. 8. Misactivation duration: measuring how long a smart device stays lit up when it is misactivated. The vertical lines show the range between the 10^{th} and the 90^{th} percentile of the observed misactivation duration (if there are more than n=10 misactivation samples), the bars show the range of activations between the 25^{th} and 75^{th} percentile, and the thick horizontal lines show the median duration. The plot is sorted by P75.

Device	Most misactivating show by time	Most misactivating show by words
google/ok-google	The Big Bang Theory (0.60)	The Big Bang Theory (0.73)
homepod/hey-siri	Greenleaf (0.58)	Narcos (1.00)
invoke/cortana	The West Wing (0.93)	The West Wing (1.04)
echo2/alexa	Gilmore Girls (0.38)	Jane The Virgin (0.34)
echo3/alexa	The L Word (0.18)	The L Word (0.27)
echo2/echo	Riverdale (0.95)	Riverdale (1.43)
echo3/echo	The West Wing (0.56)	The West Wing (0.63)
echo2/computer	The Office (U.S.) (0.64)	Narcos (0.80)
echo3/computer	Greenleaf (0.58)	Greenleaf (0.94)
echo2/amazon	The Office (U.S.) (0.56)	Narcos (1.20)
echo3/amazon	Riverdale (0.38)	Riverdale (0.57)
ALL DEVICES	The West Wing (4.26)	Narcos (6.21)

Table 3. Most misactivating shows. List of most misactivating shows by device in terms of activations per hour (second column) and activations per 10,000 words (third column).

in the much less common 90^{th} percentile case (P90), we measured a duration of 10 s in the case of Homepod, followed by 9 s in the case of Google Home Mini, and 8 s in the case of Echo Dot 2^{nd} gen. (Echo wake word).

Takeaways. Our analysis identified activations lasting 10 s (in the P90 case), which is long enough to record sensitive information from the environment. But also more common cases, such as the P75 and the median cases (respectively, 7 s and 5 s), show durations still high enough to expose some of the context of a conversation.

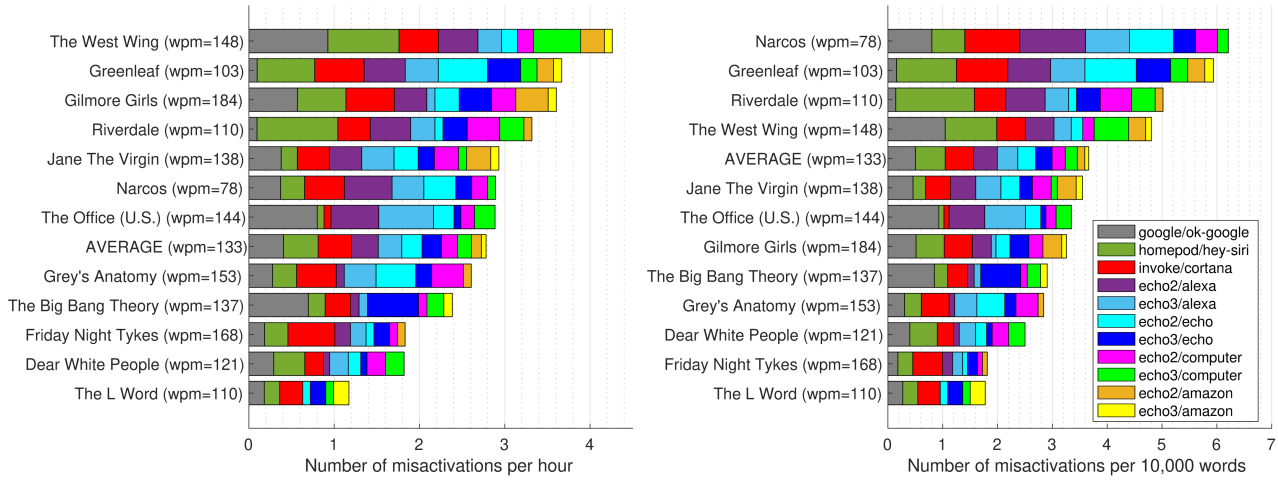
4.5 Misactivations by Shows

We now analyze whether the dialogue characteristics of the TV shows played have any impact on the number of misactivations under the assumption that users exhibiting the same dialogue characteristics are more likely to trigger misactivations (RQ5). We first answer this research question quantitatively by measuring the number of misactivations for each device and wake word with respect to the amount of content exposure in terms of time and of amount of dialogue. Then, we qualitatively analyze the results, identifying dialogue characteristics that cause the most misactivations.

The overall results are reported in Fig. 9a (playing time) and Fig. 9b (amount of dialogue). The show where we observed the most misactivations across all the speakers is The West Wing (4.26 misactivations/hour), while the show we have observed the most misactivations per word is Narcos (6.21 misactivations per 10,000 words). This difference in results is in part due to the different density of speech between the two shows (148 wpm for The West Wing *vs.* 78 wpm for Narcos, see Tab 1): the former has more dialogue per time unit, which increases the probability of such dialogue containing misactivating words; whereas the latter has the highest number of misactivating words, and therefore with the same number of words it is able to trigger the most overall misactivations. If we consider the show misactivations with respect to individual smart speakers and wake words, we see a large variability by smart speaker, playing time, and by number of words (see Tab 3). This suggests that the behavior of the devices may be affected by the type of content played.

Unfortunately we do not have the means to validate this hypothesis or to reliably infer generalizable dialogue characteristics since most misactivations are nondeterministic, but we can still see some patterns: Narcos, which tops the list in terms of misactivations per amount of dialogue, has many misactivations triggered by Spanish dialogue. Moreover, by analyzing samples of the other shows at the top of the list we have seen several instances of misactivations triggered by words not pronounced clearly, for example with a heavy accent or with a low voice.

Takeaways. We have found evidence that shows with a high amount of dialogue per time generate more misactivations and therefore that there is a correlation between the number of activations and the amount of dialogue. Also, we have found evidence that smart speakers misactivate more when they are exposed to unclear dialogue, such as a foreign language, or garbled speech. This suggests that smart speaker users who do not speak English clearly, or that are farther away from the smart speaker (lower voice volume) may have an additional risk of privacy exposure.



(a) Activations vs. show time.

(b) Misactivations vs. dialogue (words).

Fig. 9. Comparison of show: differences in the misactivation of different show, by elapsed time and by amount of dialogue. "wpm" is the number of words per minute.

4.6 Misactivations by Region

All our previous analyses focused on the US testbed. In this analysis we analyze whether the number of misactivations and the distribution of their durations in the UK testbed is similar or different with respect to the US one, thus answering the question if the region has any effects on the risk of triggering a recording from the smart speakers (RQ6). For example, devices in the UK may expect a UK accent and thus misactivate differently than US devices expecting a US accent.

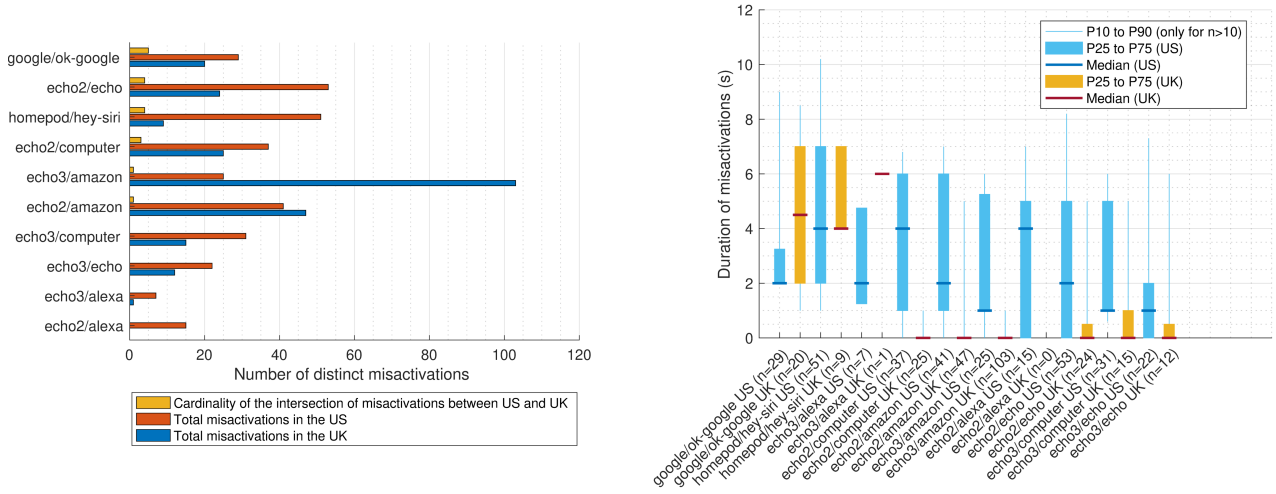
Fig. 10a shows the number of unique misactivations in each testbed, and the ones in common in the US and UK testbed. We can see that the intersection is very small, ranging from zero for most Amazon devices, to 5 misactivations in the case of Google Home Mini. This is unsurprising since we have seen that the majority of misactivations have a low repeatability and high nondeterminism among multiple experiments on the US testbed (see Fig. 5b); therefore, we would have expected something similar when comparing the two different testbeds. For the same reason, we can explain the fact that Google Home Mini has the largest number of misactivations in common due to the fact that its misactivations have the largest repeatability also within a single testbed.

One surprising result is the absolute number of misactivations that is consistently higher in the US for all devices (except for the Echo Dots with "Amazon" keyword), which shows in general a tendency of the US testbed to misactivate more. Another interesting result is how the duration of misactivations compares among the two testbeds (see Fig. 10b): the US testbed has the longest misactivations among all devices, but this may be caused by the fact that the US testbed has more misactivations, and thus more opportunities for a device to produce long misactivations.

Takeaways. To a certain degree we can explain some differences between the US and UK testbed as a result of the nondeterminism of misactivations. Another possible explanation is the presence of some differences in the test environment, which may affect sound reflections and propagations, and result in different sounds being sensed by the smart speakers under test. Moreover, devices in the UK may have their voice recognition engine trained differently (*i.e.*, for a British-English accent) or may run a different firmware. Because of these differences and our limited knowledge of device internals, we cannot confirm or deny the hypothesis that misactivations are also dependent on the region. However, we nonetheless observe strong discrepancies between our US and UK results, which motivate additional experiments to clarify if such discrepancies are caused by the different region or other differences in the testing environment.

4.7 Misactivating Words

In this last analysis we investigate RQ7, *i.e.*, to understand what false wake words typically misactivate the devices, and if there are any patterns that can explain the misactivations we observe. To answer this question, we manually analyzed the subtitles related to the misactivations that appeared at least three times between



(a) Number of misactivations (US vs. UK).

(b) Duration of misactivations (US vs. UK).

Fig. 10. Misactivation in the US testbed vs. UK testbed: measuring the misactivations differences on US devices vs. UK devices.

Words	Some patterns	Some examples from the subtitles
OK/Hey Google	Words rhyming with "Hey" or "Hi" (*e.g.*, "They" or "I"), followed by hard "G" or something containing "ol".	"okay ... to go", "maybe I don't like the cold", "they're capable of", "yeah ... good weird", "hey .. you told", "A-P ... I won't hold".
Hey Siri	Words rhyming with "Hey" or "Hi" (*e.g.*, "They" or "I"), followed by a voiceless "s"/"f"/"th" sound and a "i"/"ee" vowel.	"hey ... missy", "they ... sex, right?", "hey, Charity", "they ... secretly", "I'm sorry", "hey ... is here", "yeah. I was thinking", "Hi. Mrs. Kim", "they say ... was a sign", "hey, how you feeling".
Alexa	Sentences starting with "I" followed by a "K" or a voiceless "S".	"I care about", "I messed up", "I got something", "it feels like I'm".
Echo	Words containing a vowel plus "k" or "g" sounds.	"head coach", "he was quiet", "I got", "picking", "that cool", "pickle", "Hey, Co.".
Computer	Words starting with "comp" or rhyming with "here"/"ear".	"Comparisons", "I can't live here", "come here", "come onboard", "nuclear accident", "going camping", "what about here?".
Amazon	Sentences containing combinations of "was"/"as"/"goes"/"some" or "I'm" followed by "s", or words ending in "on/om".	"it was a", "I'm sorry", "just ... you swear you won't", "I was in", "what was off", "life goes on", "have you come as", "want some water?", "he was home".
Cortana	Words containing a "K" sound closely followed by a "R" or a "T".	"take a break ... take a", "lecture on", "quartet", "courtesy", "according to".

Table 4. Misactivating words. List of some misactivating patterns among repeatable misactivations that appear at least in three experiments (see §A for the full text of such activations). For each wake word we explain the pattern we have seen and some of the words that (loosely) conform to that pattern. This list is not exhaustive; we found hundreds of misactivations during this study.

our base experiments and the confirmatory ones. Tab. 4 shows, for each wake word, some of the activating patterns and words we have found from the subtitles of the misactivations. The full closed captions of the misactivations we consider in this analysis, including show name, episode, and timestamps, are reported in Appendix A.

What is interesting about this list is the stark difference between the actual misactivating words and the wake word for the devices. Despite all of the considered misactivations having some degree of repeatability, in many cases we could not find any clear patterns and were unable to reproduce the misactivation by repeating the closed captions with our own voice. This means that the actual speaking voice, its tone, its accent, background noise/music of the TV show, and the acoustics of the experiment environment, all play a role in producing the conditions for a misactivation.

As part of this analysis we have also searched the closed captions for each TV show played for wake words, and we found several cases where wake words occur in the closed captions, but the audio from the show does not activate the relevant speakers in any of the experiments we performed: 4 occurrences of "Echo" never lead to an activation; 4 occurrences of "Amazon" appear, but only 3 lead to an activation; 60 occurrences of "Computer" appear, but only 12 cause an activation. All the other wake words do not appear in any closed captions.

By listening to some of the misactivations with no apparent wake word similarity, we have found additional evidence that hard-to-understand content increases the risk of misactivation (for example the presence of singing, in addition to non-English dialogue).

Takeaways. From our misactivation analysis, we can say that there are cases where the misactivation is ex-

plainable (similar-sounding words to the wake words), and many where we cannot identify why the smart speaker misactivated. Fortunately, we did not find any clear evidence of consistent undocumented wake words that are malicious or completely unrelated to the real ones. Instead, on the one hand we found evidence of attempts from the manufacturers to detect some variations of their wake words, and also attempts to avoid activating when a wake word is spoken out of context (this is supported by the fact that most occurrences of the word "computer" in the closed captions did not trigger an activation). On the other hand, the fact that a smart speaker may be activated using fake wake words that can be crafted using the patterns in Tab. 4 may allow an attacker (for example, a TV commercial or a YouTube video) to activate the smart speaker, thus exposing user privacy and potentially issuing commands without the user suspecting that it was intentional.

5 Discussion

Possible causes of non-determinism. Our analysis has shown that the majority of the misactivations are not repeatable, meaning that the same piece of audio material, under the same conditions, sometimes triggers misactivations and sometimes it does not. Since the specifications of how smart speakers detect activations are not known, the best we can do is to make some educated guesses. A first source of non-determinism may be due to the analog/digital conversion of the audio material happening twice: first, when our testbed converts digital audio material to analog sounds using its attached speakers; second, when the smart speakers convert the analog sounds sensed by their microphones into the digital audio they process. Such conversions introduce stochastic noise to the system, which may cause non-deterministic differences in the audio material every time it is played. A second source of non-determinism may be due to the algorithms the devices use to recognize their wake word. For example, if a device uses a recognition model that is updated based on previous input (as we suspected for the Echo Dot devices, see §4.3), several instances of the same input may yield different classification results, which may appear non-deterministic.

Mitigations. We now mention some mitigation strategies from both a user and manufacturer perspective.

User mitigations. One approach is to use hardware features (*e.g.*, using the "mic off" toggle) or other devices to prevent smart speakers from activating at all. This includes "privacy armor" such as a bracelet [11] that emits ultrasonic frequency audio that interferes with smart speaker microphones to prevent them from detecting conversations. Another approach is to use a different wake word recognition. For instance, Snowboy [12] provides this service via open-source code, giving its users more control and transparency over activations. Like "privacy armor", this approach requires custom hardware since off-the-shelve smart speakers do not support any customization. A common theme of mitigations is that it places the onus on consumers to enable privacy-enhancing features, with varying barriers to adoption.

Manufacturer mitigations. Besides improving the quality of voice recognition, manufacturers can reduce the occurrence of misactivations by allowing different levels of wake-word sensitivity, so that users can decide the trade-off between missing real activations and misactivations. A complementary strategy is to allow more wake word customizations, so that words with a lower chance of causing misactivations can be used. Finally, since current voice assistants are already able to distinguish a real activation from a misactivation when processed in the cloud, manufacturers could implement such capability into the smart speaker itself to avoid sending misactivation recordings to the cloud at all.

Policy implications. When a smart speaker misactivates, it may transmit sounds and voices from its environment to own cloud services and record them. Since voice recordings can be considered private, sensitive information, an important issue is compliance with privacy regulations.

For example, the European General Data Privacy Regulations (GDPR) require "data protection by design" [13] and it is unclear if a device recording with no user authorization complies with that. Another important point is that several privacy regulations (including GDPR) require the user to agree to the collection of data, define what data is collected, how it is used, and who is processing it. In our context misactivations are by definition unauthorized by the user, but data is still captured and processed. Also, even if a privacy policy has been accepted by the smart speaker owner, the device can still misactivate and record other people who have never accepted the device's privacy policy.

Another example is the Illinois Biometric Information Privacy Act (BIPA) [14], which requires explicit consent for the collection of biometric identifiers such as a user voice profile. Our experiments provide evidence that some smart speakers adapt how they activate over

time (perhaps to reduce misactivations), and this may be implemented using data that are voice profiles.

Open questions. Several interesting questions are not addressed by this study and are fertile topics for future work. For example, our corpus of audio data comes from TV shows, and as such a key open question is how smart speakers react to other stimuli, such as non-verbal noises, a dictionary of words, and voices using different languages and accents. A related question is whether such stimuli can identify undocumented wake words or sounds, or reveal discriminatory biases, that could have potential privacy and policy implications. Lastly, an important question is how smart speaker ecosystems use and share the data they gather from their environments.

6 Related Work

The rapid introduction of always-on smart home assistants into households, businesses, and public spaces has raised a number of concerns from privacy advocates. While these devices offer convenient voice-based interactions, their microphones are always listening for wake words. While prior work [15] investigated whether mobile apps surreptitiously record and transmit audio from smartphones (and did not find any evidence of such behavior), the authors do not consider voice assistant misactivations, nor do they analyze smart speakers. Some researchers proposed alternative ways for controlling their activation [16, 17], while others have even gone as far as resorting to jamming devices to defend against the voice assistants [11], or introducing a filter for blocking emotions or sensitive conversations from these [18, 19].

Recently, a number of commercial IoT security tools and applications have entered the consumer market which enable device identification and network destination analysis, although they do not provide analysis of voice activations in smart speakers [20, 21]. In [22], the authors attempted remote voice command execution and session hijacking attacks on the Alexa devices. Their findings highlight some security vulnerabilities in these devices. Qualitative studies have also indicated a lack of understanding of privacy options of smart speakers and a complicated trust relationship with speaker companies [23]. Some other works analyzed privacy and security issues of smart speakers such as their network behavior [24, 25] and their skills ecosystems [26–29].

A number of researchers have carried out activation attacks on smart speakers, including injecting inaudible and invisible commands into voice assistants using lasers [30], directional amplitude-modulated ultrasound beams [31], or hidden voice commands unintelligible to human listeners [32]. In this paper we specifically focus on activations which are unintentional, or based on misclassifications under normal operational circumstances.

Compared to prior work, this is the first systematic study of the misactivations exhibited by a set of popular smart speakers from four manufacturers, using a large variety of audio materials played back at these devices in a controlled manner in two different jurisdictions.

7 Conclusion

In this paper we have measured how popular smart speakers misactivate when exposed to 134 hours of video content containing more than one million words of dialogue. Our findings include some good news: we did not find evidence of malicious or intentional misactivations, meaning that the misactivations we observed may just be caused by a suboptimal wake word recognition engine. There is also some bad news: the fact that misactivations may be unintentional does not mitigate the risk of privacy exposure since conversations and other audio captured as a result of misactivations are often sent over the Internet and stored on remote servers.

As smart speakers become increasingly pervasive in everyday life, there is an urgent need to understand the behavior of this ecosystem and its impacts on consumers. This work represents a first analysis of smart speaker behavior at scale. As part of future work, we intend to further investigate this popular ecosystem in terms of how audio data is used and shared by voice assistant providers, what are additional privacy and security risks from interacting with these devices, and how can we better protect consumers in this environment.

To support further research, all software and data we produced as part of this work are publicly available at `https://moniotrlab.ccis.neu.edu/smart-speakers-study`.

8 Acknowledgments

We thank the anonymous reviewers and our shepherd Aziz Mohaisen for their constructive feedback. The research in this paper was partially supported by the NSF (CNS-1909020) and the EPSRC (Databox EP/N028260/1, DADA EP/R03351X/1, and HDI EP/R045178/1).

References

[1] Amazon, *Alexa voice assistant. Accessed on 02/28/2020*, https://en.wikipedia.org/wiki/Amazon_Alexa.

[2] Google, *Google Assistant. Accessed on 02/28/2020*, https://assistant.google.com.

[3] Apple, *Siri voice assistant. Accessed on 02/28/2020*, https://www.apple.com/siri/.

[4] Cortana, *Cortana voice assistant. Accessed on 02/28/2020*, https://www.microsoft.com/windows/cortana.

[5] Forrester, *Smart Home Devices Forecast, 2017 To 2022 (US). Accessed on 02/28/2020*, https://www.forrester.com/report/Forrester+Data+Smart+Home+Devices+Forecast+2017+To+2022+US/-/E-RES140374.

[6] Artem Russakovskii, *Google is permanently nerfing all Home Minis because mine spied on everything I said 24/7. Accessed on 02/28/2020*, https://www.androidpolice.com/2017/10/10/google-nerfing-home-minis-mine-spied-everything-said-247/.

[7] VRT NWS, *Google employees are eavesdropping, even in your living room. Accessed on 02/28/2020*, https://www.vrt.be/vrtnws/en/2019/07/10/google-employees-are-eavesdropping-even-in-flemish-living-rooms/.

[8] G. Fowler, *Alexa has been eavesdropping on you this whole time. Accessed on 02/28/2020*, https://www.washingtonpost.com/technology/2019/05/06/alexa-has-been-eavesdropping-you-this-whole-time/.

[9] Amazon, *Alexa Cloud Documentation. Accessed on 02/28/2020*, https://www.amazon.com/gp/help/customer/display.html?nodeId=GHXNJNLTRWCTBBGW.

[10] Google, *Google Cloud Documentation. Accessed on 02/28/2020*, https://support.google.com/websearch/answer/6030020?co=GENIE.Platform%3DDesktop&hl=en.

[11] Y. Chen, H. Li, S.-Y. Teng, S. Nagels, Z. Li, P. Lopes, B. Zhao, and H. Zheng, "Wearable Microphone Jamming," in *Conference on Human Factors in Computing Systems 2020 (CHI '20)*, 2020.

[12] KITT.AI, *Snowboy, a hotword detection engine. Accessed on 02/28/2020*, http://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=3004&ChapterID=57.

[13] EU Parliament, *General Data Protection Regulation (GDPR). Accessed on 02/28/2020*, https://gdpr-info.eu/.

[14] Illinois General Assembly, *Biometric Information Privacy Act (BIPA). Accessed on 02/28/2020*, http://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=3004&ChapterID=57.

[15] E. Pan, J. Ren, M. Lindorfer, C. Wilson, and D. R. Choffnes, "Panoptispy: Characterizing Audio and Video Exfiltration from Android Applications," in *Privacy Enhancing Technologies Symposium (PETs '18)*, 2018.

[16] A. Mhaidli, M. Venkatesh, Y. Zou, and F. Schaub, "Listen Only When Spoken To: Interpersonal Communication Cues as Smart Speaker Privacy Controls," in *Privacy Enhancing Technologies Symposium (PETs '20*, 2020.

[17] B. Karmann, *Project Alias. Accessed on 02/28/2020*, https://bjoernkarmann.dk/project_alias.

[18] C. Champion, I. Olade, C. Papangelis, H. Liang, and C. Fleming, "The smart speaker blocker: An open-source privacy filter for connected home speakers," *arXiv preprint arXiv:1901.04879*, 2019.

[19] R. Aloufi, H. Haddadi, and D. Boyle, "Privacy preserving speech analysis using emotion filtering at the edge," in *17th Conference on Embedded Networked Sensor Systems (SenSys '19)*, 2019, pp. 426–427.

[20] J. Ren, D. J. Dubois, D. Choffnes, A. M. Mandalari, R. Kolcun, and H. Haddadi, "Information Exposure for Consumer IoT Devices: A Multidimensional, Network-Informed Measurement Approach," in *Proc. of the Internet Measurement Conference (IMC '19)*, 2019.

[21] D. Y. Huang, N. Apthorpe, G. Acar, F. Li, and N. Feamster, "IoT Inspector: Crowdsourcing Labeled Network Traffic from Smart Home Devices at Scale," *arXiv preprint arXiv:1909.09848*, 2019.

[22] I. Castell-Uroz, X. Marrugat-Plaza, J. Solé-Pareta, and P. Barlet-Ros, "A first look into alexa's interaction security," in *15th ACM Intern.l Conf. on Emerging Networking EXperiments and Technologies (CoNEXT '19)*, 2019.

[23] J. Lau, B. Zimmerman, and F. Schaub, "Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-Seeking Behaviors with Smart Speakers," *Proceedings of the ACM on Human-Computer Interaction (issue CSCW)*, vol. 2, no. 1, pp. 1–31, 2018.

[24] S. Kennedy, H. Li, C. Wang, H. Liu, B. Wang, and W. Sun, "I Can Hear Your Alexa: Voice Command Fingerprinting on Smart Home Speakers," in *2019 IEEE Conference on Communications and Network Security (CNS '19)*, June 2019, pp. 232–240.

[25] N. Apthorpe, D. Reisman, S. Sundaresan, A. Narayanan, and N. Feamster, "Spying on the smart home: Privacy attacks and defenses on encrypted iot traffic," *arXiv preprint arXiv:1708.05044*, 2017.

[26] D. Kumar, R. Paccagnella, P. Murley, E. Hennenfent, J. Mason, A. Bates, and M. Bailey, "Skill Squatting Attacks on Amazon Alexa," in *27th USENIX Security Symposium (USENIX Security '18)*, Aug. 2018, pp. 33–47.

[27] A. Alhadlaq, J. Tang, M. Almaymoni, and A. Korolova, "Privacy in the Amazon Alexa skills ecosystem," in *10th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs '17)*, 2017.

[28] N. Zhang, X. Mi, X. Feng, X. Wang, Y. Tian, and F. Qian, "Understanding and mitigating the security risks of voice-controlled third-party skills on amazon alexa and google home," *arXiv preprint arXiv:1805.01525*, 2018.

[29] R. Mitev, M. Miettinen, and A.-R. Sadeghi, "Alexa Lied to Me: Skill-based Man-in-the-Middle Attacks on Virtual Assistants," in *2019 ACM Asia Conf. on Computer and Communications Security (ASIACCS '19)*, 2019, pp. 465–478.

[30] T. Sugawara, B. Cyr, S. Rampazzi, D. Genkin, and K. Fu, *Light Commands: Laser-Based Audio Injection on Voice-Controllable Systems. Accessed on 02/28/2020*, https://lightcommands.com/.

[31] R. Iijima, S. Minami, Y. Zhou, T. Takehisa, T. Takahashi, Y. Oikawa, and T. Mori, "Audio Hotspot Attack: An Attack on Voice Assistance Systems Using Directional Sound Beams and its Feasibility," *IEEE Transactions on Emerging Topics in Computing*, 2019.

[32] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, "Hidden voice commands," in *25th USENIX Conference on Security Symposium (USENIX Security '16)*. USENIX Association, 2016.

A Closed Captions from Repeatable Activations

In this appendix we report, for each device and wake word, the timestamped closed captions for all the misactivations that appear in three or more experiments. This is the data we used to infer the patterns reported on Tab. 4.

We selected closed captions starting 5 s before the beginning of a misactivation, and lasting a total of 6 s. We decided these numbers based on manually watching a handful of samples for each smart speaker and wake word combination.

From the samples we watched, we can confirm that the closed captions are properly synched and that they capture the part of dialogue lighting up the device. Each closed caption has two timestamps at the beginning, the one where the closed caption is supposed to appear on the video and the one where it is supposed to disappear.

We believe this information, together with the software and complete dataset we released[1], to be valuable for researchers, device manufacturers, and regulators to support and reproduce our findings.

A.1 OK/Hey Google (Home Mini)

```
[S12.Ep7] The Big Bang Theory - The Grant Allocation Derivation
652 655 Someone's making decisions.
656 658 I'm reviewing these proposals.
658 660 Yeah. 'Cause you're the boss man,
660 661 telling people what's what.
661 664 I like it.

[S12.Ep16] The Big Bang Theory - The D & D Vortex
1073 1076 Okay, great. Thanks. Bye.
1076 1078 Okay, where were we?
1081 1085 I was about to go all Wrath of Khan on the ogres.

[S12.Ep19] The Big Bang Theory - The Inspiration Deprivation
455 457 What do you say? We could both use a break.
457 459 Come on, I'll do it with you.
459 460 Okay, but not in the same tank.
460 462 I already shared a uterus with my twin sister.
462 464 I don't need to go through that again.

[S4.Ep7] Friday Night Tykes - Ain't Gotta Cheat Us to Beat Us
1515 1517 AND THEN WE'VE GOT THE PLAYOFFS.
1517 1518 YOU KNOW, WE DON'T PLAYER HATE.
1518 1519 WE'RE GONNA KEEP GOING, FULL THROTTLE,
1519 1521 AND JUST KEEP GOING.
1521 1524 WE CAN'T LET THIS GAME DEFINE OUR SEASON.
1524 1525 KEEP MOVING. IT'S JUST A GAME.

[S5.Ep10] The L Word - Lifecycle
2135 2136 THANK YOU.
2136 2137 WILL YOU COME VISIT ME
2137 2139 WHEN I GO BACK TO SCHOOL?
2139 2140 MAYBE.
2140 2143 I DON'T LIKE THE COLD, THOUGH.
2143 2145 I'LL KEEP YOU WARM.

[S6.Ep7] The L Word - Last Couple Standing
1555 1557 OKAY, DO YOU THINK YOU NEED TO GO OVER IT AGAIN?
1557 1558 I THINK I'M GOOD.
1558 1560 DO YOU THINK, OR ARE YOU SURE?
1560 1561 BECAUSE YOU KNOW WHAT, THEY'VE GOT WIGS AND SPANDEX.
1561 1562 THIS IS NO FUCKING JOKE.
1562 1564 WE HAVE NO IDEA WHAT THEY'RE CAPABLE OF.

[S5.Ep9] Jane The Virgin - Chapter Ninety
1077 1079 Ooh!
1079 1082 Things are getting pretty steamy between you two.
1082 1083 [LAUGHS] Hard to wait?

[S7.Ep14] Gilmore Girls - Farewell, My Pet
771 773 HAROLD, I'VE ALREADY PAID THE BILL.
773 774 AND THIS IS THE ROOM RATE.
774 776 TIMES THREE NIGHTS. YEP.
776 778 OKAY, AND WHAT IS THIS CHARGE FOR, EXACTLY,
778 779 UNDER ROOM SERVICE?
779 781 THAT'S...FOR THE ROOM SERVICE THAT YOU ORDERED.

[S7.Ep14] Gilmore Girls - Farewell, My Pet
2274 2278 YEAH, I GUESS I'M NOT EITHER. IT'S WEIRD.
2278 2279 BUT GOOD WEIRD?
2279 2281 GREAT WEIRD.

[S7.Ep15] Gilmore Girls - I'm a Kayak, Hear Me Roar
1404 1407 Now, that's not true. He's made
          you feel incompetent, too.
1407 1410 I guess Logan was excited that his dad
          wanted to take us out
1410 1412 so that's sweet.
1412 1415 Hey, have you told grandma and grandpa
          about you and dad yet?

[V2.Ep5] Dear White People - Chapter V
537 539 Let me guess. You...
541 542 A-P girl.
542 544 Mm. Obvi.
544 545 I won't hold it against you.
```

A.2 Hey Siri (Homepod)

```
[S11.Ep24] The Big Bang Theory - The Bow Tie Asymmetry
272 274 Hey, guys, look who I have.
275 279 -Hey, guys. Hey, Shelly.
        -I'm so glad you made it, Missy.
279 282 This is my fiancée, Amy. Amy, this is my sister.

[S15.Ep16] Grey's Anatomy - Blood and Water
499 502 talk about this later.
502 505 Please, um... continue.
505 506 Nico: Okay. Well, the first step would be
506 508 to make an incision in the thigh.

[S15.Ep17] Grey's Anatomy - And Dream of Sheep
311 313 Babies die in womb all the time.
313 314 Papa, you're working very fast.
314 316 Y-You are too impatient. Hey!
316 318 You don't like the way I work, you can get out.
318 321 I'm not saying that I -- Go!

[S15.Ep17] Grey's Anatomy - And Dream of Sheep
2025 2031 ♫♫
2031 2033 ♪ What I need ♪
2033 2036 [ Crying, sniffling ]
2036 2038 ♪ Is you ♪

[S3.Ep11] Greenleaf - The End Is Near
2174 2177 It's more than I can get from her.
2177 2180 But they are the weaker sex, right?
2183 2185 Basie...

[S3.Ep2] Greenleaf - The Space Between
2097 2100 Well, all right then.
2102 2104 Hey, Charity.

[S3.Ep3] Narcos - Follow the Money
```

1 https://moniotrlab.ccis.neu.edu/smart-speakers-study

499 501 [Peña in English] It was true.
501 503 While the terms of surrender were being finalized,
504 506 the brothers had gone underground to avoid capture.
507 510 They abandoned their swank estates,
secretly moving to new homes,

[S3.Ep10] Friday Night Tykes - Babies from Texas
113 115 THIS WAS ONE OF THE LAST YEARS TOGETHER.
115 117 A LOT OF THE KIDS WILL BE PLAYIN' JUNIOR HIGH BALL
NEXT YEAR,
118 119 AND SO IT'S GONNA BE DIFFERENT.
119 125 ♪♪

[S3.Ep10] Friday Night Tykes - Babies from Texas
2038 2040 IF YOU GOT SOME BABIES FROM LONG BEACH
2040 2042 ON THE FOOTBALL FIELD, MAKE SOME NOISE.
2042 2045 [cheers and applause]
2045 2047 - BE SEEIN' YA, HOMIE. - WE GONNA EAT, ALL RIGHT?

[S6.Ep4] The L Word - Leaving Los Angeles
2202 2204 I'M SORRY.

[S6.Ep6] The L Word - Lactose Intolerant
1739 1742 YOU'RE WELCOME.
1745 1747 HEY, T, YOUR CAR IS HERE!
1747 1749 I'M COMING!
1749 1751 I JUST...

[S5.Ep15] Jane The Virgin - Chapter Ninety-Six
294 295 And lover.
295 297 Okay. Yeah.
298 299 It's good, right?
299 300 I think it's good.
300 302 I also feel strange saying that,
302 304 like, maybe I'm delusional.

[S7.Ep11] Gilmore Girls - Santa's Secret Stuff
111 114 OH, HELLO. OKAY. OH.
114 116 OH. WELL, ALL RIGHT. [LAUGHS]
116 118 SO, YOU MADE IT HERE OKAY? YEAH.
118 120 I WAS THINKING -- ALL THAT TIME IN ENGLAND,
120 122 YOU MIGHT FORGET WHICH SIDE OF THE ROAD TO DRIVE ON.

[S7.Ep13] Gilmore Girls - I'd Rather Be in Philadelphia
555 557 In a way, it's their fault that Richard's here.
557 559 Mom, what do you mean?
559 560 Two and a half months ago, I read an article
560 562 that said fish has been shown to prevent heart attacks
562 565 and stroke, and has innumerable other health benefits.

[S7.Ep15] Gilmore Girls - I'm a Kayak, Hear Me Roar
63 65 - Driving? - Yeah, driving.
65 67 - Mom, what's going on? - Want to go for a drive?
67 69 Um, sure. Let's go for a drive.

[S7.Ep16] Gilmore Girls - Will You Be My Lorelai Gilmore
662 664 You buy them, and then you take them home.
664 666 What if they don't fit next to the bed?
666 668 Then you'll get a new bed.
668 670 - Hi, Mrs. Kim. - Lorelai.
670 672 - How's business? - People die, go bankrupt.

[S7.Ep14] The West Wing - Two Weeks Out
636 637 HAND FEEL BETTER IN THAT THING?
637 640 THEY SAY REPLACING SHEILA WAS A SIGN OF WEAKNESS.
641 642 INSIDE BASEBALL.

[S7.Ep15] The West Wing - Welcome to Wherever You Are
749 752 HELLO? HELLO?
752 754 HEY, HOW YOU FEELING?
754 756 TERRIBLE-- THEY JUST HUNG UP ON ME.
756 757 AP? REUTERS.
757 758 SNOTTY. LITTLE BIT.

[S7.Ep18] The West Wing - Requiem
934 936 BLACKJACK IS RISKY, STOCK-PICKING IS RISKY...
936 939 YEAH, WELL, I'M MORE OF A CREDIT UNION KIND OF GUY.
939 941 ANY SCENARIO LOSES US ALLIES.
941 943 IF YOU CAN CONCEIVABLY WORK WITH SELLNER...

A.3 Alexa (Echo Dot 2^{nd} gen.)

[S5.Ep16] Jane The Virgin - Chapter Ninety-Seven
142 146 Because I wanted to end the book on an uplifting note.
146 148 Joy, love, all that stuff.
148 151 Yeah, I get that, but I don't 100 percent care.
151 153 I care about what the book needs.

[S7.Ep18] Gilmore Girls - Hay Bale Maze
2370 2371 Yeah, I'm, uh, I'm sorry, too.
2371 2374 - No, no, no. Let me go first. - 'Okay.'
2374 2376 I messed up.

[S7.Ep16] The West Wing - Election Day
1340 1342 THEN WE'LL FIND SOMETHING ELSE TO DO.
1342 1343 WHAT DO WE GOT?
1343 1346 I GOT YOU SOMETHING.
1346 1348 A GIFT? WHATEVER.

A.4 Alexa (Echo Dot 3^{rd} gen.)

[S7.Ep11] The West Wing - Internal Displacement
121 124 EXCEPT IT'S COLD AND DARK. WHAT?
124 125 I WAS MAKING A JOKE.
125 127 OH, YOU DON'T HAVE TO DO THAT.
127 128 RELAXING MAKES ME NERVOUS.
128 130 IT FEELS LIKE I'M MISSING SOMETHING.
130 132 YOU WANT A DRINK? NO.

A.5 Echo (Echo Dot 2^{nd} gen.)

[S3.Ep3] Narcos - Follow the Money
770 772 The dimmer is over here.

[S4.Ep7] Friday Night Tykes - Ain't Gotta Cheat Us to Beat Us
1026 1027 I WILL BE THE HEAD COACH FOR THE JUNIORS
1027 1030 JYSF VENOM FOOTBALL TEAM NEXT YEAR.
1031 1033 AND HEAD COACH CHRIS DAVIS IS NOT GONNA MESS AROUND.

[S3.Ep8] Riverdale - Chapter Forty-Three - 'Outbreak'
429 432 Hey, I got my GED now.
432 434 Hmm? I started this whole place up.
434 437 I am a legit businesswoman now.

[S3.Ep11] Riverdale - Chapter Forty-Six - 'The Red Dahlia'
596 599 Betty was being a gnat as usual, T.T.
599 601 And I'm afraid I don't have any patience for it.
601 604 Back off, Betty. It's been a rough day.

[S3.Ep12] Riverdale - Chapter Forty-Seven - 'Bizarrodale'
1351 1354 I did it. I told my dad.
1354 1355 What? You did? How'd it go?
1355 1358 He was quiet and weird at first,

[S3.Ep13] Riverdale - Chapter Forty-Eight - 'Requiem for a
Welterweight'
121 124 Sorry, you're what?
124 126 Edgar says I'm finally ready.
126 130 All of the women from the Farm are gathering Sunday
night at the new facility.

[S3.Ep18] Riverdale - Chapter Fifty-Three - 'Jawbreaker'
1983 1984 Jason
1985 1986 or you.
1987 1990 Cheryl, Jason's a ghost.

[S7.Ep8] The West Wing - Undecideds
1182 1185 YOU KNOW HOW HOT IT IS THERE? YEAH.
1185 1187 WE HAD ABOUT ENOUGH OF PICKING WITH THE COTTON.
1187 1188 HOW'D IT GO, SIR?
1188 1191 A LOT LIKE IT DID THE LAST 80 TIMES.

[V1.Ep10] Dear White People - Chapter X
136 137 if that's cool.
140 141 Sam? That cool?

[V3.Ep8] Dear White People - Chapter VIII
85 87 Professor Brown's... pickle?
89 90 Sorry, I'm usually more artful than this.
90 92 Oh, are you?

[V3.Ep9] Dear White People - Chapter IX
137 138 Hey, Co.
138 139 [Coco] Hey, you look nice.
139 140 Whoa.
141 143 -Is everything okay? -What do you mean?

A.6 Echo (Echo Dot 3^{rd} gen.)

[S3.Ep18] Riverdale - Chapter Fifty-Three - 'Jawbreaker'
1983 1984 Jason
1985 1986 or you.
1987 1990 Cheryl, Jason's a ghost.

[S7.Ep8] The West Wing - Undecideds
1180 1182 WHO DO YOU KNOW THAT WANTED AN AVOCADO-PICKING JOB?
1182 1185 YOU KNOW HOW HOT IT IS THERE? YEAH.
1185 1187 WE HAD ABOUT ENOUGH OF PICKING WITH THE COTTON.
1187 1188 HOW'D IT GO, SIR?
1188 1191 A LOT LIKE IT DID THE LAST 80 TIMES.

[V1.Ep10] Dear White People - Chapter X
145 147 Sounds great. Um, I'll be right back.
149 153 Is there a button I'm supposed to push or something?

A.7 Computer (Echo Dot 2^{nd} gen.)

[S12.Ep22] The Big Bang Theory - The Maternal Conclusion
1181 1183 I love you.
1183 1185 I love you, too.
1187 1190 Now both of you say it to me.

[S15.Ep12] Grey's Anatomy - Girlfriend in a Coma
148 150 [Telephone rings in distance]
150 151 [Sighs]
151 153 You had the husband? Yeah.
153 155 Ice skate to the tibia. You?
155 157 Ouch. Wife had a hell of a concussion.

[S15.Ep17] Grey's Anatomy - And Dream of Sheep
172 174 [Telephone rings in distance]
174 175 So you're saying I shouldn't cancel it?
175 176 I'm saying it shoots tomorrow,
176 178 and it's too late to cancel.
178 179 Don't be nervous.
179 180 You're gonna do great.
180 182 Okay. [Chuckling] Okay?

[S2.Ep16] Greenleaf - The Pearl
327 328 Both of you.
329 330 Thank you.
330 333 Thank you. They're beautiful.
333 336 Oh. [CAR HORN HONKS]

[S3.Ep8] Greenleaf - Dea Abscondita
381 384 Well, my heart's ticking just fine. Thank you.
384 386 Well, you're a better man than me.
386 389 Comparisons are odious.
389 391 But I won't disagree.

[S3.Ep9] Greenleaf - Runaway Train
1264 1266 Sophia!
1267 1270 Well, what'd you think?
1270 1272 Your spirit didn't hear that?
1272 1274 I don't think I have a spirit.

[S3.Ep9] Narcos - Todos Los Hombres del Presidente
717 718 Don't worry.
722 725 Don't worry. Everything will be okay.

[S3.Ep9] Riverdale - Chapter Forty-Four - 'No Exit'
1326 1328 It's your turn, Red.

[S3.Ep12] Riverdale - Chapter Forty-Seven - 'Bizarrodale'
2274 2277 Moose, your friends are here, and I'm here,
2277 2279 -and you can live with any of us. -Kevin...
2279 2281 I...
2282 2284 I can't live here.

[S3.Ep20] Riverdale - Chapter Fifty-Five - 'Prom Night'
1749 1753 A few months back, I found something out.
1753 1754 Something I should have told you.
1756 1757 You're gonna want to sit down for this.

[S5.Ep7] Jane The Virgin - Chapter Eighty-Eight
580 583 But you know what?
583 586 It's a new day. It is.
586 588 And a beautiful one at that.
589 591 I'm glad you think so.

[S5.Ep17] Jane The Virgin - Chapter Ninety-Eight
1122 1124 when you saw him, your heart would glow.
1124 1126 Didn't exactly go down like pickle juice.
1126 1128 LATIN LOVER NARRATOR: Very little does.
1128 1129 I get it.
1129 1131 That was a long time ago.

[S5.Ep19] Jane The Virgin - Chapter One Hundred
129 131 Okay.
134 137 We're...moving to New York.
137 138 What?
138 140 LATIN LOVER NARRATOR: So, yeah.

[S7.Ep11] The West Wing - Internal Displacement
532 535 SQUEEZE HIM IN. REALLY?
535 537 WHY NOT?
537 539 YOU I NEED. COME HERE. WHAT'S WRONG?
539 541 CLOSE THE DOOR. I DIDN'T DO IT.
541 543 CLOSE THE DOOR. TOBY DID IT.

[S7.Ep16] The West Wing - Election Day
219 221 DID YOU EVER... "COME ONBOARD?"

[S7.Ep20] The West Wing - The Last Hurrah
788 790 HEY, BOB, YOU GOT THE EXIT POLLS?
790 792 RIGHT HERE. PRETTY SIMPLE.
792 796 WE LOST NEVADA BY 70,000 VOTES BECAUSE OF THE
NUCLEAR ACCIDENT.

[S9.Ep20] The Office (U.S.) - Paper Airplane
668 672 ONE OF YOU WILL WALK AWAY WITH $2,000.
672 674 - YEAH!
674 677 - OKAY, HERE YOU HAVE JUST KNOCKED OVER THE BEAKER.
677 679 THE CHEMICALS SPLASHED IN YOUR EYE.

[S9.Ep21] The Office (U.S.) - Livin' the Dream
2158 2160 - TENTS?
2160 2162 ARE YOU THINKING OF GOING CAMPING?
2162 2164 I THOUGHT YOU FOUND NATURE VULGAR.

[S9.Ep23] The Office (U.S.) - Finale
1688 1692 YOU CAN TAKE THAT TO THE BANK.
1692 1695 - YOU READY? - [chuckles] YOU KIDDING?
1695 1696 I WAS BORN READY.
1696 1698 [mimicking heavy metal guitars]

A.8 Computer (Echo Dot 3^{rd} gen.)

[S15.Ep18] Grey's Anatomy - Add It Up
1979 1981 No, good. We're good.
1981 1982 Hey, hey, give me one second, okay? [Door opens]
1982 1984 Yeah.
1984 1985 [Door closes]
1985 1987 So...
1987 1989 I-I'm curious.

[S15.Ep24] Grey's Anatomy - Drawn to the Blood
641 644 That's a nasty business.
644 645 Thanks, Altman. I needed this.
645 647 You needed a punctured rectum?
647 649 [Chuckling] I needed a distraction

649 652 from whatever's going on in the conference room.

[S3.Ep9] Greenleaf - Runaway Train
1264 1266 Sophia!
1267 1270 Well, what'd you think?
1270 1272 Your spirit didn't hear that?
1272 1274 I don't think I have a spirit.

[S3.Ep10] Greenleaf - The Promised Land
2224 2225 What about her?
2226 2229 And...
2229 2231 be sure you know what you're talking about

[S2.Ep9] Narcos - Nuestra Finca
464 465 Hello.
468 470 And to what do I owe this lovely surprise?
470 473 We're at the warehouse in Manrique.

[S2.Ep9] Narcos - Nuestra Finca
2609 2614 Where is the money? The money we gave you.
I need it back.
2614 2617 We spent it. It's gone.
2617 2620 Relax. I'm not going to hurt you.

[S3.Ep9] Narcos - Todos Los Hombres del Presidente
717 718 Don't worry.
722 725 Don't worry. Everything will be okay.

[S9.Ep17] The Office (U.S.) - The Farm
247 250 I'M ON STEP EIGHT OF ALCOHOLICS ANONYMOUS,
250 252 STEP NINE OF NARCOTICS ANONYMOUS.
252 254 I'M HERE TO MAKE AMENDS.
254 257 I'VE BEEN HARD TO DEAL WITH OVER THE PAST YEARS.

[S9.Ep20] The Office (U.S.) - Paper Airplane
668 672 ONE OF YOU WILL WALK AWAY WITH $2,000.
672 674 - YEAH!
674 677 - OKAY, HERE YOU HAVE JUST KNOCKED OVER THE BEAKER.
677 679 THE CHEMICALS SPLASHED IN YOUR EYE.

[V2.Ep7] Dear White People - Chapter VII
515 518 Last night was so rough,
518 519 I couldn't even fuck it away.
519 520 Hmm.
522 523 What matters is you keep trying.

A.9 Amazon (Echo Dot 2^{nd} gen.)

[S12.Ep6] The Big Bang Theory - The Imitation Perturbation
591 594 I'm just surprised you don't remember our first kiss.
594 597 (sighs) Fine. It was on Halloween.
597 598 Are you agreeing just to shut me up?
598 600 You got another way? I'm all ears.

[S3.Ep5] Greenleaf - Closing Doors
2420 2422 ...I'm sorry.

[S3.Ep10] Greenleaf - The Promised Land
1879 1882 Stop lying to me, Jacob! I'm not lying, okay?
1882 1884 It was just a kiss.
1884 1887 And it was a nothing kiss, because I shut it down.
1887 1890 I told her I don't play like that anymore.

[S3.Ep2] Narcos - The Cali KGB
517 518 There he is.
518 521 Javi Peña, el jefe.
522 525 Never thought I'd see the day, but I'm sure as shit
glad to see you now.
525 527 -Duff, it's been a long time. -Mm-hmm.

[S3.Ep7] Narcos - Sin Salida
945 947 Paola. Please.
949 951 The Americans are going to try to arrest Miguel.
951 954 I'm going to make sure they succeed.

[S3.Ep9] Narcos - Todos Los Hombres del Presidente
369 372 -And we capture him while they're moving. -Exactly.
373 374 How?
375 376 Leave that to me.
377 381 I'll stay here in Bogotá. You coordinate the operation
in Cali.

[S3.Ep14] Riverdale - Chapter Forty-Nine - 'Fire Walk With Me'
553 554 and I was lucky enough
554 557 to have friends who helped me get through that time.
557 559 Just... You swear
559 561 you won't call Social Services.

[S3.Ep14] Riverdale - Chapter Forty-Nine - 'Fire Walk With Me'
665 667 and was squatting at the gym. That's how I found him.
667 669 Sounds like me, sophomore year.
669 671 Guys, he has a branding on his arm.
671 675 The same one the Warden gave me when I was in juvie.
It says he a sacrifice.

[S3.Ep20] Riverdale - Chapter Fifty-Five - 'Prom Night'
662 665 uh, actually, we'll take two, please.
665 667 Well, well, well.
667 669 What was off is now back on again.
669 671 You owe me a cherry phosphate.

[S5.Ep8] Jane The Virgin - Chapter Eighty-Nine
2470 2472 [PHONE BUZZES]
2472 2475 Hey, my favorite person in the world. Are you
almost home?
2475 2477 Ugh, no, still at the laundromat.
2477 2480 LATIN LOVER NARRATOR: He definitely has dirty laundry.

[S5.Ep9] Jane The Virgin - Chapter Ninety
1767 1769 [IN SPANISH] I know he would.
1769 1771 I was just thinking about how different
1771 1773 this wedding will be from my first.
1775 1777 When it was just me and your grandfather,

[S5.Ep15] Jane The Virgin - Chapter Ninety-Six
831 833 I can get off work early and pick up the girls
833 835 so you don't have to.
835 837 No, no, no, life goes on,
837 840 and I still have plenty of agents to hear from.
840 844 And me and the girls are making real progress.

[S7.Ep12] Gilmore Girls - To Whom It May Concern
1015 1017 Oh, I like your office. It's cozy.
1017 1020 Hmm. That's one way of describing it.
1020 1022 So have you come as a loving granddaughter
1022 1025 visiting your grandfather or as an obsequious student

[S7.Ep9] The West Wing - The Wedding
530 532 YOU WOULD HAVE LIKED HIM.
532 534 I'M GETTING TOO OLD FOR THIS JOB.
534 536 REALLY.
536 539 IT WAS THEIR RELATIVES, WHO WE DON'T PARTICULARLY KNOW,
539 541 AND OUR RELATIVES, WHO WE DON'T PARTICULARLY LIKE.

[S7.Ep16] The West Wing - Election Day
1435 1437 YOU WANT SOME WATER?
1437 1440 YOU HAVE A BOTTLE OVER THERE?
1440 1441 TAP WATER.

[S9.Ep16] The Office (U.S.) - Moving On
1815 1817 - SO THERE'S NO MARKETING DEPARTMENT?
1817 1819 both: NO.
1819 1821 - YOU KNOW, TIMES WERE TOUGH.
1821 1822 I WAS UNEMPLOYED.
1822 1824 I WAS STILL HEARTBROKEN OVER YOU.

[S9.Ep20] The Office (U.S.) - Paper Airplane
85 88 - "BE CAREFUL OF THAT BEAKER. IT CONTAINS DANGEROUS ACID!"
88 89 - IT DOES NOT SAY "DANGEROUS."
89 91 AND THERE'S NO EXCLAMATION POINT.
91 94 - WELL, I'M JUST--I'M TRYING TO BRING SOME LIFE TO IT.
94 96 LAST WEEK I GOT AN AGENT.

[S9.Ep23] The Office (U.S.) - Finale
442 443 HE USED TO CALL IT A KELEVEN.
443 445 HE TOLD DWIGHT, "A MISTAKE PLUS KELEVEN
445 447 GETS YOU HOME BY SEVEN."
447 449 HE WAS HOME BY 4:45 THAT DAY

A.10 Amazon (Echo Dot 3^{rd} gen.)

```
[S3.Ep7] Narcos - Sin Salida
142 144 My name is Pacho Herrera.
146 149 I want Gerda Salazar and her fucking sons delivered
        to me.
150 154 Until that happens, nobody in the North Valley is safe.

[S3.Ep14] Riverdale - Chapter Forty-Nine - 'Fire Walk With Me'
665 667 and was squatting at the gym. That's how I found him.
667 669 Sounds like me, sophomore year.
669 671 Guys, he has a branding on his arm.
671 675 The same one the Warden gave me when I was in juvie.
        It says he a sacrifice.

[S5.Ep8] Jane The Virgin - Chapter Eighty-Nine
2470 2472 [PHONE BUZZES]
2472 2475 Hey, my favorite person in the world. Are you
          almost home?
2475 2477 Ugh, no, still at the laundromat.
2477 2480 LATIN LOVER NARRATOR: He definitely has dirty laundry.

[S9.Ep23] The Office (U.S.) - Finale
442 443 HE USED TO CALL IT A KELEVEN.
443 445 HE TOLD DWIGHT, "A MISTAKE PLUS KELEVEN
445 447 GETS YOU HOME BY SEVEN."
447 449 HE WAS HOME BY 4:45 THAT DAY.
```

A.11 Cortana (Invoke)

```
[S15.Ep14] Grey's Anatomy - I Want a New Drug
1536 1538 Dr. Shepherd?
1538 1542 Are you okay?
1542 1545 Do you want to take a break, take a walk?
1545 1548 I knew him.

[S7.Ep9] Gilmore Girls - Knit, People, Knit
350 352 It's not because they don't have black ties.
352 353 'Suit yourself.'
353 355 Now, what do you think
355 356 a string quartet, or something more fun
356 357 like a swing band?

[S7.Ep15] Gilmore Girls - I'm a Kayak, Hear Me Roar
2482 2483 - What are you up to today? - Today?
2484 2485 I'm going to attend a D.A.R. lecture
2486 2487 on Native American artwork.
2487 2489 Then I have a lunch with Sarah Montgomery Brown
2489 2491 and Melissa Seria and of course

[S7.Ep8] The West Wing - Undecideds
2412 2414 WE'RE TIRED OF WAITING.
2414 2416 WE'RE TIRED OF TRYING TO FIGURE OUT
2416 2419 WHY OUR CHILDREN ARE NOT SAFE.
2419 2422 AND WHY OUR EFFORTS TO MAKE THEM SAFE SEEM TO FAIL.

[S7.Ep21] The West Wing - Institutional Memory
947 949 TO THE SANTOS PEOPLE?
949 951 OH. THEY REALLY SHOULDN'T NEED REFERENCES.
951 952 THEY'RE OFFERING ME A JOB, PURELY AS A COURTESY,
952 955 WHICH I'LL PRETEND TO CONSIDER, PURELY AS A FORMALITY.
955 957 I THINK THERE ARE SOME MISTAKES

[S8.Ep21] The Office (U.S.) - Angry Andy
544 546 -WASHINGTON MONUMENT. -OKAY.
548 550 -EIFFEL TOWER. -OKAY, OKAY.
550 552 [mouse clicking]
552 553 HMM.

[S9.Ep18] The Office (U.S.) - Promos
742 744 "SO YOU'RE AN IDIOT. AND I AM HAWT,
744 747 "ACCORDING TO PEOPLE ON THIS SITE WHO HAVE A BRAIN.
747 750 NEVER COMMENT ON THIS PAGE EVER AGAIN."
750 752 "HE IS HAWT."
```