Exploring Accidental Triggers of Smart Speakers

Lea Schönherr, Maximilian Golla†, Thorsten Eisenhofer, Jan Wiele, Dorothea Kolossa, Thorsten Holz⋆
Ruhr University Bochum
† *Max Planck Institute for Security and Privacy*
⋆ *CISPA Helmholtz Center for Information Security*

Abstract

Voice assistants like Amazon's Alexa, Google's Assistant, Tencent's Xiaowei, or Apple's Siri, have become the primary (voice) interface in smart speakers that can be found in millions of households. For privacy reasons, these speakers analyze every sound in their environment for their respective *wake word* like "Alexa," "Jiǔsì'èr líng," or "Hey Siri," before uploading the audio stream to the cloud for further processing. Previous work reported on examples of an inaccurate wake word detection, which can be tricked using similar words or sounds like "cocaine noodles" instead of "OK Google."

In this paper, we perform a comprehensive analysis of such *accidental triggers*, i. e., sounds that should not have triggered the voice assistant, but did. More specifically, we automate the process of finding accidental triggers and measure their prevalence across 11 smart speakers from 8 different manufacturers using everyday media such as TV shows, news, and other kinds of audio datasets. To systematically detect accidental triggers, we describe a method to artificially craft such triggers using a pronouncing dictionary and a weighted, phone-based Levenshtein distance. In total, we have found hundreds of accidental triggers. Moreover, we explore potential gender and language biases and analyze the reproducibility. Finally, we discuss the resulting privacy implications of accidental triggers and explore countermeasures to reduce and limit their impact on users' privacy. To foster additional research on these sounds that mislead machine learning models, we publish a dataset of more than 350 verified triggers as a research artifact.

1 Introduction

In the past few years, we have observed a huge growth in the popularity of voice assistants, especially in the form of smart speakers. Most major technology companies, among them Amazon, Baidu, Google, Apple, Tencent, and Xiaomi, have developed an assistant. Amazon is among the most popular brands on the market: the company reported in 2019 that it had sold more than 100 million devices with *Alexa* on board; there were more than 150 products that support this voice assistant (e. g., smart speakers, soundbars, headphones, etc.) [10]. Especially smart speakers are on their way of becoming a pervasive technology, with several security and privacy implications due to the way these devices operate: they continuously analyze every sound in their environment in an attempt to recognize a so-called *wake word* such as "Alexa," "Echo," "Hey Siri," or "Xiǎo dù xiǎo dù." If and only if a wake word is detected, the device starts to record the sound and uploads it to a remote server, where it is transcribed, and the detected word sequence is interpreted as a command. This mode of operation is mainly used due to privacy concerns, as the recording of all (potentially private) communication and processing this data in the cloud would be too invasive. Furthermore, the limited computing power and storage on the speaker prohibits a full analysis on the device itself. Hence, the recorded sound is sent to the cloud for analysis once a wake word is detected.

Unfortunately, the precise detection of wake words is a challenging task with a typical trade-off between usability and security: manufacturers aim for a low false acceptance and false rejection rate [56], which promotes a certain wiggle room for an adversary. As a result, it happens that these smart speakers trigger even if the wake word has not been uttered. First exploratory work on the error patterns of voice-driven user input has been done by Vaidya et al. [69]. In their 2015 paper, the authors explain how Google's voice assistant, running on a smartphone, *misinterprets* "cocaine noodles" as "OK Google" and they describe a way to exploit this behavior to execute unauthorized commands such as sending a text, calling a number, or opening a website. Later, Kumar et al. [41] presented an attack, called *skill squatting*, that leverages transcription errors of a list of words sounding similar to existing Alexa skills. Their attack exploits the *imperfect transcription* of the words by the Amazon API and routes users to malicious skills with similar-sounding names. A similar attack, in which the adversary exploits the way a skill is invoked, has been described by Zhang et al. [75].

Such research results utilize instances of what we call an *accidental trigger*: a sound that a voice assistant mistakes for its wake word. Privacy-wise, this can be fatal, as it will induce the voice assistant to start a recording and stream it to the cloud. Inadvertent triggering of smart speakers and the resulting accidentally captured conversations are seen by many as a privacy threat [14, 21, 46]. When the media reported in summer 2019 that employees of the manufacturer listen to voice recordings to transcribe and annotate them, this led to an uproar [18, 71]. As a result, many companies paused these programs and no longer manually analyze the recordings [23, 33, 43].

In this paper, we perform a systematic and comprehensive analysis of accidental triggers to understand and elucidate this phenomenon in detail. To this end, we propose and implement an automated approach for systematically evaluating the resistance of smart speakers to such accidental triggers. We base this evaluation on candidate triggers carefully crafted from a pronouncing dictionary with a novel phonetic distance measure, as well as on available AV media content and bring it to bear on a range of current smart speakers. More specifically, in a first step, we analyze vendor's protection mechanisms such as cloud-based wake word verification systems and acoustic fingerprints, used to limit the impact of accidental triggers. We carefully evaluate how a diverse set of 11 smart speakers from 8 manufacturers behaves in a simulated living-room-like scenario with different sound sources (e. g., TV shows, news, and professional audio datasets). We explore the feasibility of artificially crafting accidental triggers using a pronouncing dictionary and a weighted, phone-based Levenshtein distance metric and benchmark the robustness of the smart speakers against such crafted accidental triggers. We found that a distance measure that considers phone-dependent weights is more successful in describing potential accidental triggers. Based on this measure, we crafted 1-, 2-, and 3-grams as potential accidental triggers, using a TTS service and were able to find accidental triggers for all tested smart speakers in a fully automated way.

Finally, we give recommendations and discuss countermeasures to reduce the number of accidental triggers or limit their impact on users' privacy.

To summarize, we make the following key contributions:

1. By reverse-engineering the communication channel of an Amazon Echo, we are able to provide novel insights on how commercial companies deal with such problematic triggers in practice.

2. We develop a fully automated measurement setup that enables us to perform an extensive study of the prevalence of accidental triggers for 11 smart speakers from 8 manufacturers. We analyze a diverse set of audio sources, explore potential gender and language biases, and analyze the identified triggers' reproducibility.

3. We introduce a method to synthesize accidental triggers with the help of a pronouncing dictionary and a weighted phone-based Levenshtein distance metric. We demonstrate that this method enables us to find new accidental triggers in a systematic way and argue that this method can benchmark the robustness of smart speakers.

4. We publish a dataset of more than 350 accidental triggers to foster future research on this topic.[1]

2 Understanding Accidental Triggers

In this section, we provide the required background on wake word detection. Furthermore, we describe how Amazon deals with accidental triggers and how we analyzed and reverse engineered an Amazon Echo speaker. Finally, we provide an overview of smart speaker privacy settings. In general, accidental triggers are the consequence of the trade-off between specificity and sensitivity, namely the false rejection and the false acceptance rate. Whenever a wake word recognizer is trained, the system aims to minimize both of these errors.

2.1 Wake Word Recognition

To enable natural communication between the user and the device, automatic speech recognition (ASR) systems built into smart speakers rely on a far-field voice-based activation. In contrast to a push-to-talk model, where speech recognition is only active after a physical button is pressed, smart speakers *continuously* record their surroundings to allow hands-free use. After detecting a specific *wake word*, also known as hotword or keyword, the smart speaker starts to respond. The wake word recognition system is often a lightweight DNN-based ASR system, limited to a few designated words [6, 29, 73]. All these systems use a small DNN, e.g., in the form of a TDNN, to map the audio input into a representation that describes the likelihoods of all units (e.g. a phone) of the wake word in each frame. Based on this DNN output, the system decides on the presence of the wake word, for which purpose different strategies may be used. Either a graph search is conducted through the keyword's phone sequences given the DNN output [6, 29] or the likelihoods of the phones are combined to a score. To guarantee its responsiveness, the recognition runs locally and is therefore limited by the computational power and storage of the speaker. For example, we found the *Computer* wake word model for an Amazon Echo speaker to be less than 2 MB in size, running on a 1GHz ARM Cortex-A8 processor. The speaker uses about 50% of its CPU time for the wake word recognition process. In addition to the wake word, the model also detects a stop signal ("Stop") to interrupt the currently running request. Especially when used in environments with ambient noise

[1] They are available at https://unacceptable-privacy.github.io, where we also provide example videos.

from external sources such as TVs, a low false acceptance and false rejection rate is much harder to achieve for these systems [56].

The device will only transmit data to the respective server *after* the wake word has been recognized locally. Hence, activating the wake word by an accidental trigger will lead to the upload of potentially sensitive and private audio data, and should, therefore, be avoided as far as possible.

In some cases, a speaker misinterprets another word or sound as its wake word. If the misinterpreted word is unrelated to the configured wake word, we refer to this event as an *accidental trigger*. To limit the consequences of such false wakes, vendors started to augment the local wake word recognition with a *cloud-based wake word verification*. Moreover, there is an *acoustic fingerprint*-based mechanism in place that prevents a speaker from triggering when listening to certain audio sequences observed in TV commercials and similar audio sources. We describe both of these mechanisms in more detail in Section 2.4.

2.2 Voice Profiles, Sensitivity, and Audible Feedback

Voice profiles, also referred to as "Voice Match" or "Recognize My Voice" feature, are a convenience component of modern voice assistants [72]. The requisite voice training was introduced with iOS 9 (2015), and Android 8 (2017) to build context around questions and deliver personalized results. On smartphones, a voice profile helps to recognize the user better [7]. Vendors explain that without a profile, queries are simply considered to be coming from guests and thus will not include personal results [26].

In contrast to voice assistants on phones, smart speakers are intended to be activated by third parties, such as friends and visitors. Thus, voice profiles do not influence whether a smart speaker is activated or not when the wake word is recognized. In shared, multi-user environments, voice profiles enable voice assistants to tell users apart and deliver personalized search results, music playlists, and communication. The feature is also not meant for security, as a similar voice or recording can trick the system [24]. In our experiments, voice profiles were not enabled or used.

In April 2020, Google introduced a new feature that allows users to adjust the wake word's responsiveness to limit the number of accidental activations [25]. In our experiments, we used the "Default" sensitivity. In a short test, we found that this "Default [0]" setting caused 30 accidental wake ups, while choosing the most private setting, called "Least Sensitive [-2]," resulted in 28 (-7%), thus only reducing the number of accidental triggers by 2. In contrast, the least private mode, called "Most Sensitive [+2]," responded to 44 (+47%) times, showing that this new feature is more useful for increasing the responsiveness "particularly in a noisy environment," [27] and has almost no positive impact on users' privacy.

Another potential privacy protection feature is to play a chime, sound effect, or another audible tone when the wake word is detected, to notify the user about the potential (mis-) activation. On smartphones, this audible feedback is enabled by default. However, smart speakers do not play a sound, but instead activate their LED activity indicator as a visual cue. For people with visual impairments, some smart speakers include an accessibility setting to enable the playback of such an auditory feedback [28].

2.3 Alexa Internals

In the following, we describe how we analyzed and reverse engineered an Amazon Echo speaker (1st Gen.). The speaker was bought in February 2017 and was equipped with firmware version 647 588 720 from October 2019.

Rooting To obtain root access on an Amazon Echo speaker, we follow a method described by Clinton et al. [16] that was later refined by Barnes [9]. To decrypt and analyze the speaker's communication with the Alexa API, we inject a shared object that dumps all negotiated pre-master secrets into a file, which we later use to decrypt the TLS protected traffic recorded in the PCAP files using the tool *Wireshark*.

From the Wake Word to the Response Echo's ASR engine is called *Pryon* and started from a fork of the open-source speech recognition toolkit Kaldi [55]. There are four wake words models, i. e., *Alexa*, *Computer*, *Echo*, and *Amazon*, divided by the two different device types *doppler* and *pancake* (Echo and Echo Dot), and four different languages/regions (en-US, es-US, en-GB, and de-DE). The local automatic speech recognition daemon, ASRD, uses Pryon to detect the configured wake word. The ASRD represents its certainty for recognizing the wake word with a *classifier score* between 0.0 and 1.0. The communication between the Echo and Amazon's cloud relies on the latency-optimized SPDY protocol.

The following six steps are relevant: ① In normal use, a wake word score above 0.57 is categorized as an "Accept." A score between 0.1 (notification threshold) and 0.57 will be categorized as a "NearMiss." The threshold for an accept is lowered to 0.43, if the device is playing music. A near miss will not trigger the LED indicator, and no audio will be processed or uploaded to the Amazon cloud. ② In contrast, an "Accept" will activate the encoding of the currently recorded audio. The LED indicator turns on and starts to indicate the estimated direction of the speech source. Moreover, the ASRD informs the cloud about an upcoming audio stream, together with the information where in the stream the ASRD believes to have recognized the wake word. ③ The cloud then runs its own detection of the wake word (cf. Section 2.4.1). ④ If the wake word is recognized, the cloud will send a transcription of the audio back to the device, e. g., "what are the Simpsons." In response, Echo will switch the LED indicator to a blue

circulating animation. In the meantime, the conversational intelligence engine in the cloud tries to answer the question. ⑤ Next, the cloud will respond with the spoken answer, encoded as an MP3 file. Echo then notifies the cloud that it is playing the answer, and the LED indicator switches to a blue fade in/out animation. At the same time, the cloud requests the device to stop uploading the microphone input. ⑥ When the AlexaSpeechPlayer has finished, the ASRD informs the cloud about the successful playback, and Echo switches the LED indicator off.

Summarizing, we can confirm that the examined device is only transmitting microphone input to Amazon's cloud if the LED indicator is active and hence acting as a trustworthy indicator for the user. Based on a packet flow analysis, this is also true for all other voice assistants. One exception is the smart speaker built by Xiaomi, which seems to upload speech that can be considered a near miss to overrule the local ASR engine, without switching on the LED indicator.

2.4 Reducing Accidental Triggers

Next, we focus on two methods that vendors deploy to prevent or recover from accidental triggers.

2.4.1 Cloud-Based Wake Word Verification

As mentioned before, the local speech recognition engine is limited by the speaker's resources. Thus, in May 2017, Amazon deployed a two-stage system [36], where a low-power ASR on the Echo is supported by a more powerful ASR engine in the cloud. A few months later, Apple described a similar cloud-based verification system, where the main speech recognizer "sends a cancellation signal" if it detects something other than "Hey Siri" [6].

Accordingly, accidental triggers can be divided into two categories: (i) *local triggers* that overcome the local classifier, but get rejected by the cloud-based ASR engine, and (ii) *local + cloud triggers* that overcome both. While a local trigger switches the LED indicator on, a subsequent question "*{accidental local trigger}*, will it rain today?" will not be answered. In cases where the cloud does not confirm the wake word's presence, it sends a command to the Echo to stop the audio stream. Surprisingly, the entire process from the local recognition of the wake word to the moment where Echo stops the stream and switches off the LED indicator only takes about $1-2$ seconds. In our tests, we observe that during this process, Echo uploads at least $1-2$ seconds of voice data, approx. 0.5 seconds of audio before the detected wake word occurs, plus the time required to utter the wake word (approx. another second). In cases where the cloud-based ASR system also detects the wake word's presence, the accidental trigger can easily result in the upload of 10 or more seconds of voice data. During our experiments, we found that all major smart speaker vendors use a cloud-based verification system, including Amazon, Apple, Google, and Microsoft.

2.4.2 Acoustic Fingerprints

To prevent TV commercials and similar audio sources from triggering Echo devices, Amazon uses an acoustic fingerprinting technique. Before the device starts to stream the microphone input to the cloud, a local database of fingerprints is evaluated. In the cloud, the audio is checked against a larger set of fingerprints. The size of the local database on an Amazon Echo (1st Gen.) speaker is limited by its CPU power and contains 50 entries. This database gets updated approximately every week with 40 new fingerprints, which mostly contain currently airing advertisements [57] for Amazon products. Until mid-May 2020, the database contained dates and clear text descriptions of the entries. Since then, only hash values are stored. The database still contained 9 fingerprints from 2017, e. g., the "Echo Show: Shopping Lists" YouTube video.

We evaluate some of the entries by searching for the commercials and find videos where the ASRD reports a "[...] strong fingerprint match." However, we also observed false positives, where fingerprint matches against commercials were found, which were not present in the database, leaving the robustness of the technique [30] in question. We found that the metricsCollector process on the Echo speaker periodically collects and uploads the detected fingerprints. This is particularly concerning for privacy since it shows an interest of Amazon in these local fingerprint matches that could easily be combined with the cloud matches and be abused to build viewing profiles of the users [49]. If the wake word is spoken on live TV, Amazon will register a large peak in concurrent triggers with the same audio fingerprint and automatically request the devices to stop [57].

2.5 Smart Speaker Privacy Settings

To learn more about how vendors handle their users' data, we requested the voice assistant interaction history from Amazon, Apple, Google, and Microsoft using their respective web forms. Among the tested vendors, Apple is the only manufacturer that does not provide access to the voice data but allows users to request their complete deletion.

Table 1: Smart Speaker Privacy Settings

	Voice Recordings				Local
Vendor	**Opt-Out**	**Retention**	**Delete**	**Report**	**Trigger**
Amazon	Yes	3, 18 months	A, R, I	Yes	Yes
Apple	Yes	6, 24 months	A	-	-
Google	Yes	3, 18 months	A, R, I	No	Yes
Microsoft	Yes*	Unspecified	A, I	No	No

*Cannot speak to Cortana anymore; A=All, R=Range, I=Individual.

In Table 1, we analyze whether a user is able to opt-out of the automatic storing of their voice data, how long the recordings will be retained, the possibility to request the deletion of the recordings, and whether recordings can be reported as problematic. Furthermore, we checked if false activations through accidental triggers, i. e., local triggers, are visible to the user ("Audio was not intended for Alexa"). Apple reports storing the voice recordings using a device-generated random identifier for up to 24 months but promises to disassociate the recordings from related request data (location, contact details, and app data) [34], after six months. In contrast, customers of Amazon and Google can choose between two different voice data retention options. According to Google, the two time frames of 3 and 18-months are owed to *recency* and *seasonality* [54]. Microsoft's retention policy is more vague, but they promise to comply with legal obligations and to only store the voice data "as long as necessary."

3 Evaluation Setup

In this section, we describe our evaluated smart speakers and the datasets we used for our measurement study.

3.1 Evaluated Smart Speakers

In our experiments, we evaluate 11 smart speakers as listed in Table 2. The smart speakers have been selected based on their market shares and availability [45, 60]. In the following, with the term *smart speaker*, we refer to the hardware component. At the same time, we use the term *voice assistant* to refer to cloud-assisted ASR and the conversational intelligence built into the speaker.

Since its introduction in 2014, the Amazon Echo is one of the most popular speakers. It enables users to choose between four different wake words ("Alexa," "Computer," "Echo," and "Amazon"). In our experiments, we used four Echo Dot (3rd Gen.) and configured each to a different wake word. Similarly, for the Google Assistant, we used a Home Mini speaker, which listens to the wake words "OK Google" and "Hey Google." From Apple, we evaluated a HomePod speaker with "Hey Siri" as its wake word. To test Microsoft's Cortana, we bought the official Invoke smart speaker developed by Harman Kardon that recognizes "Cortana" and "Hey Cortana."

Moreover, we expanded the set by including non-English (US) speaking assistants from Europe and Asia. We bought three Standard Chinese (ZH) and one German (DE) speaking smart speaker. The Xiaomi speaker listens to "Xiǎo ài tóngxué" (小爱同学), which literately translates to "little classmate." The Tencent speaker listens to "Jiǔsì'èr líng" (九四二零), which literately translates to the digit sequence 9-4-2-0. The wake word is a phonetic replacement of "Jiùshì ài nǐ," which translates to "just love you." The Baidu speaker listens to "Xiǎo dù xiǎo dù" (小度小度), which literately translates to "small degree," but is related to the smart device product line Xiaodu (little "du" as in Bai*du*). Finally, we ordered the Magenta Speaker from the German telecommunications operator Deutsche Telekom, which listens to "Hallo," "Hey," and "Hi Magenta." In this case, magenta refers to a product line and also represents the company's primary brand color. Deutsche Telekom has not developed the voice assistant in-house. Instead, they chose to integrate a third-party white-label solution developed by SoundHound [38]. While the speaker also allows accessing Amazon Alexa, we have not enabled this feature for our measurements. The Magenta Speaker is technically identical to the Djingo speaker [51], which was co-developed by the French operator Orange.

3.2 Evaluated Datasets

In the following, we provide an overview of the datasets used to evaluate the prevalence of accidental triggers. We included media to resemble content, which is likely played in a typical US household to simulate an environment with ambient noise from external sources such as TVs [56]. Moreover, we considered professional audio datasets used by the machine learning community.

TV Shows The first category of media is TV shows. We considered a variety of different genres to be most representative. Our list comprises popular shows from the last 10 years and includes animated series and a family sitcom, a fantasy drama, and a political thriller. Our English (US) TV show dataset includes *Game of Thrones*, *House of Cards*, *Modern Family*, *New Girl*, and *The Simpsons*.

News The second category is newscasts. As newscasts tend to be repetitive, we used one broadcast per day and television network only. The analyzed time frame covers news broadcast between August and October 2019. Our English (US) newscasts dataset includes *ABC World News*, *CBS Evening News*, *NBC Nightly News*, and *PBS NewsHour*.

Professional Datasets The third category is professional audio datasets. Due to the costly process of collecting appropriate training datasets and the accessibility of extensive and well-analyzed datasets, we considered professional datasets commonly used by the speech recognition community.

- *LibriSpeech* [52]: An audio dataset created by volunteers who read and record public domain texts to create audiobooks. It contains 1,000 hours of speech. The corpus has been built in 2015 and is publicly available; it is a widely used benchmark for automatic speech recognition.
- *Mozilla Common Voice* [22]: The dataset is based on an ongoing crowdsourcing project headed by Mozilla to create a free speech database. At the time of writing,

Table 2: Evaluated Smart Speakers

ID	Assistant	Release	Wake Word(s)	Lang.†	Smart Speaker	SW. Version
VA1	Amazon: Alexa	2014	*Alexa*	en_us, de_de	Amazon: Echo Dot (v3)	392 657 0628
VA2	Amazon: Alexa	2014	*Computer*	en_us, de_de	Amazon: Echo Dot (v3)	392 657 0628
VA3	Amazon: Alexa	2014	*Echo*	en_us, de_de	Amazon: Echo Dot (v3)	392 657 0628
VA4	Amazon: Alexa	2014	*Amazon*	en_us, de_de	Amazon: Echo Dot (v3)	392 657 0628
VA5	Google: Assistant	2012	*OK/Hey Google*	en_us, de_de	Google: Home Mini	191 160
VA6	Apple: Siri	2011	*Hey Siri*	en_us, de_de	Apple: HomePod	13.4.8
VA7	Microsoft: Cortana	2014	*Hey/- Cortana*	en_us	Harman Kardon: Invoke	11.1842.0
VA8	Xiaomi: Xiao AI	2017	*Xiǎo ài tóngxué*	zh_cn	Xiaomi: Mi AI Speaker	1.58.4
VA9	Tencent: Xiaowei	2017	*Jiǔsì'èr líng*	zh_cn	Tencent: Tīngtīng TS-T1	3.5.0.025
VA10	Baidu: DuerOS	2015	*Xiǎo dù xiǎo dù*	zh_cn	Baidu: NV6101 (1C)	1.34.5
VA11	SoundHound: Houndify	2015	*Hallo/Hey/Hi Magenta*	de_de	Deutsche Telekom: Magenta Speaker	1.1.2

†: In our experiments, we only considered English (US), German (DE), and Standard Chinese (ZH).

the project includes a collection of 48 languages. Our English (US) version of the dataset contains 1,200 hours of speech and has been downloaded in August 2019. As neither the environment nor the equipment for the audio recordings is controlled, the quality of the recordings differs widely.

- *Wall Street Journal* [53]: A corpus developed to support research on large-vocabulary, continuous speech recognition systems containing read English text. The dataset was recorded in 1993 in a controlled environment and comprises 400 hours of speech.

- *CHiME* [8]: The CHiME (Computational Hearing in Multisource Environments) dataset is intended to train models to recognize speech from recordings made by distant microphones in noisy environments. The *5th* CHiME challenge dataset includes recordings from a group dinner of four participants each, with two acting as hosts and two as guests. Audio signals were recorded at 20 parties, each in a different home, via six Kinect microphone arrays and four binaural microphone pairs. This dataset thus provides multi-channel recordings of highly realistic, distant-talking speech with natural background noise. In total, the dataset consists of 50 hours of recording time.

Noise We used noise recordings as a special category to test the sensitivity of the voice assistants against audio data other than speech. For this purpose, we used the noise partition of the *MUSAN* dataset [65], containing approximately 6 hours of many kinds of environmental noise (excluding speech and music).

Non-English Media To test for linguistic differences, e. g., biases between different languages, we tested one Standard Chinese (ZH) and four German (DE) TV shows. We analyzed the Chinese TV show *All Is Well* and the German-dubbed version of the TV show *Modern Family* for easy comparison. Additionally, we tested the German-dubbed version of *The Big Bang Theory*, as well as *Polizeiruf 110* and *Tatort* as examples for undubbed German TV shows. Moreover, we evaluated three shorter (each 12 hours) samples of the Chinese newscast *CCTV Xinwen Lianbo* and the German newscasts *ARD Tagesschau* and *ZDF Heute Journal*.

Female vs. Male Speakers To explore potential gender biases in accidental triggers of voice assistants, we also included two sets of randomly chosen voice data from the *LibriSpeech* dataset. Every set consisted of a female and a male 24 hour sample. Every sample was built from multiple 20-minute sequences, which themselves were made of 100 different 12 seconds audio snippets.

4 Prevalence of Accidental Triggers

Based on the datasets described above, we now explore the prevalence of accidental triggers in various media such as TV shows, newscasts, and professional audio datasets.

4.1 Approach

We start by describing our technical setup to measure the prevalence of accidental triggers across 11 smart speakers (cf. Section 3.1) using 24-hour samples of various datasets (cf. Section 3.2). The basic idea is to simulate a common living-room-like scenario, where a smart speaker is in close proximity to an external audio source like a TV.

Testing a living-room scenario is of particular interest, as it simulates a very popular smart speaker environment (43–75% of speakers are located in the living room [39]) and is also used by smart-speaker vendors when evaluating the performance of their wake-word models (cf. [2, 56, 62, 63]).

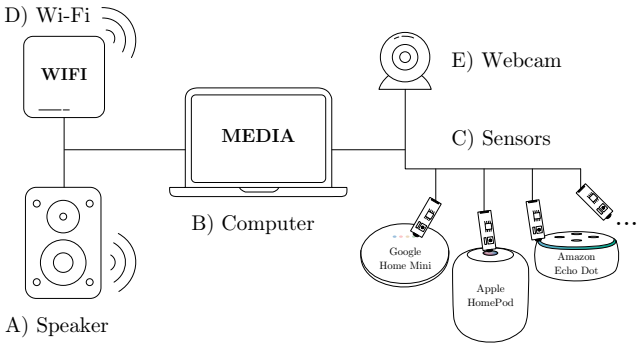

Figure 1: Setup: A loudspeaker (A) is playing media files from a computer (B). The LED activity indicators of a group of smart speakers are monitored using light sensors (C). All speakers are connected to the Internet over Wi-Fi (D). A webcam (E) is used to record a video of each measurement.

4.1.1 Measurement Setup

Hardware The measurement setup consists of five components, as depicted in Figure 1. To rule out any external interference, all experiments are conducted in a sound-proof chamber. We positioned 11 smart speakers at a distance of approx. 1 meter to a loudspeaker (A) and play media files from a computer (B). To detect any activity of the smart speakers, we attach photoresistors (C) (i. e., light sensors) on the LED activity indicator of each speaker, as one can see in Figure 2. In the case of any voice assistant activity, the light sensor detects the quick change in brightness and emits a signal to the computer (B). To prevent interference from external light sources, the photoresistors are covered by a small snippet of reusable adhesive tape.

Figure 2: Photoresistor attached to the LED indicator of a smart speaker. The sensitivity of the sensor can be adjusted via a potentiometer. Any activity is recognized and logged.

All smart speakers are connected to the Internet using a WiFi network (D). During all measurements, we record network traces using `tcpdump` to be able to analyze their activity on a network level. To verify the measurement results, we record a video of each measurement via a webcam with a built-in microphone (E). The entire setup is connected to a network-controllable power socket that we use to power cycle the speakers in case of failures or non-responsiveness.

Software To verify the responsiveness of the measurement setup, we periodically play a test signal, which consists of the wake word (e. g., "Alexa") and the stop word (e. g., "Stop") of each voice assistant (in its configured language) and a small pause between them. Overall, the test signal for all 11 speakers is approximately 2m 30s long. During the measurements, we verify that each voice assistant triggers to its respective test signal. In the case of no response, multiple or prolonged responses, all voice assistants are automatically rebooted and rechecked. As a side effect, the test signal ensures that each assistant stops any previous activity (like playing music or telling a joke) that might have been accidentally triggered by a previous run. Using this setup, we obtain a highly reliable and fully automated accidental trigger search system.

4.1.2 Trigger Detection

The process of measuring the prevalence of accidental triggers consists of three parts, as depicted in Figure 3. First, a 24-hour search is executed twice per dataset. Second, a tenfold verification of a *potential* trigger is done to confirm the existence of the trigger and measure its reproducibility. Third, a manual classification of *verified* triggers is performed to ensure the absence of the wake word or related words. In the following, we describe these steps in more detail.

I. Search: In a first step, we prepare a 24-hour audio sample consisting of multiple episodes/broadcasts of approximately 20 minutes each (with slightly different lengths depending on the source material) for each of the datasets introduced in Section 3.2. We play each of the 24-hour samples twice and log any smart speaker LED activity as an indicator for a potential trigger. The logfile includes a timestamp, the currently played media, the playback progress in seconds, and the triggered smart speaker's name. We played each audio file twice due to some changes in results that were observed when we played the same sample multiple times. These changes do not come as a great surprise, given that they are due to the internal framing of the recorded audio. Therefore, each time one plays the same audio file, the system will get a slightly different signal, with slightly shifted windows and possibly small changes in the additive noise that cannot be fully prevented but is (strongly) damped in our test environment. Also, there may be further indeterminacies up in the chain of trigger processing, as was also noted by others [21, 41].

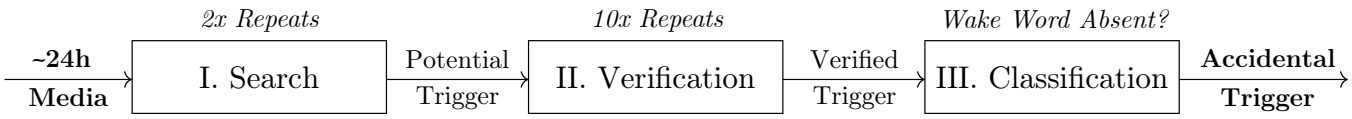

Figure 3: Trigger Detection Workflow: Every approx. 24-hour dataset is played twice. Subsequently, the existence of every *potential* trigger is confirmed. Finally, every *verified* trigger is classified as *accidental*, if the wake word or a related word is not present in the identified scene.

II. Verification: In a second step, we extract a list of *potential* triggers from the logfile and verify these triggers by replaying a 10-second snippet containing the identified scene. From the potential trigger location within the media, i. e., the playback progress when the trigger occurred, we rewind 7 seconds and replay the scene until 3 seconds after the documented trigger location. This playback is repeated ten times to confirm the existence and to measure the reproducibility of the trigger.

III. Classification : In a third step, every verified trigger is classified by reviewing a 30-second snippet of the webcam recording at the time of the trigger. Here, two independently working reviewers need to confirm the accidental trigger by verifying the correct wake word's absence. If a trigger is caused by the respective wake word or a related word such as Alexander ("Alexa"), computerized ("Computer"), echoing ("Echo"), Amazonian ("Amazon"), etc., we discard the trigger and exclude it from further analysis. Where available, the analysis is assisted by the transcriptions/subtitles of the respective dataset.

To determine the approximate distribution between *local* and *cloud*-based triggers, we expand our classification step. Instead of only determining the mere presence of the wake word or a related word, two members of our team also classify the triggers into local or local+cloud triggers. As noted in Section 2.5, not all smart speaker vendors provide access or report local triggers in their voice assistant interaction history. Thus, we use the internal processes, especially the LED timings and patterns, to classify triggers. The heuristic for that classification is based on the time the LED indicator of the speaker remains on. Based on preliminary tests, we choose a conservative threshold of 2 seconds of speaker activity to classify the trigger as local+cloud. Moreover, we use voice responses and certain LED patterns as obvious signals for a local+cloud trigger. The inter-rater reliability between our reviewers, measured by Cohen's kappa, is $\kappa \geq 0.89$ across all evaluated datasets.

4.2 Results

An overview of our results can be found in Table 3. We report the absolute counts of observed accidental triggers and actual instances of spoken wake words.

Prevalence The average number of uttered words vary across datasets. Utilizing their subtitles, we exemplarily counted the number of words in the different TV shows in Table 3. The TV show with the fewest words in 24 hours was Game of Thrones, using 126,941 words (88 wpm). In contrast, 24h of New Girl used 234,460 words (163 wpm). Across all TV shows, VA7 Hey Cortana had the highest number of accidental triggers, with one accidental trigger every 9,738 words or 1h 16min. For VA1 Alexa, we observed one accidental trigger every 29,528 words or 3h 52min. VA6 Hey Siri triggered much more rarely, with only one misactivation every 457,684 words or 60h of watching TV.

Comparison Across Speakers Looking at the four VA1-4 Amazon Echo wake words, we can see that "Amazon" (67) and "Echo" (43) trigger less often than "Alexa" (100) and "Computer" (80). Moreover, we observe that the VA5 Google Home (23) and the VA6 Apple HomePod (9) seem to be the most robust speakers of all English (US) speakers across all played datasets, and we discuss potential reasons for that in Section 7.2. Another noteworthy observation is that VA7 Microsoft Cortana triggered far more often (198) than the other speakers across all kinds of audio data.

From a qualitative perspective, the identified triggers are often confusions with similar-sounding words or sequences, such as, "a lesson" (Alexa), "commuter" (Computer), "OK, cool" (OK Google), "a city" (Hey Siri). Another category are proper names that are unknown or likely infrequently included in the training data. Examples include names of persons and states such as "Peter" and "Utah" (Computer), "Eddard" (Echo), "Montana" (Cortana), but also uncommon old English phrases such as "Alas!" (Alexa). Finally, we observed a few cases of triggers that include fictional language (*Dothraki*) or unintelligible language (*gibberish*) and two occasions of *non-speech* accidental triggers: A ringing phone triggering "Amazon" in the TV show *New Girl* and a honk made by a car horn triggering "Alexa" in the TV show *The Simpsons*.

Comparison Across Datasets When comparing across datasets, one must keep in mind that the total playback time differs across categories. While every dataset (i. e., every row in the table) consisted of 24 hours of audio data, the number of datasets per category differs. For easier comparison with

Table 3: Prevalence of Accidental Triggers and Wake Words.

			Alexa		Computer		Echo		Amazon		Ok Google		Hey Siri		Hey Cortana		Xiǎo ài tóngxué		Jiǔsì' èr líng		Xiǎo dù xiǎo dù		Hallo Magenta	
			A	W	A	W	A	W	A	W	A	W	A	W	A	W	A	W	A	W	A	W	A	W
			en_us		en_us		en_us		en_us		en_us		en_us		en_us		zh_cn		zh_cn		zh_cn		de_de	
		Tot.	**31**	**0**	**31**	**6**	**18**	**2**	**38**	**2**	**3**	**0**	**2**	**0**	**94**	**0**	**1**	**0**	**7**	**0**	**0**	**0**	**3**	**0**
		ph	**0.258**	**0.0**	**0.258**	**0.050**	**0.150**	**0.017**	**0.317**	**0.017**	**0.025**	**0.0**	**0.017**	**0.0**	**0.783**	**0.0**	**0.008**	**0.0**	**0.058**	**0.0**	**0.0**	**0.0**	**0.025**	**0.0**
TV Shows	GoT	24h	6	-	6	-	5	-	3	-	-	-	-	-	14	-	1	-	1	-	-	-	1	-
	H. o. Cards	24h	2	-	11	-	2	2	15	-	-	-	1	-	14	-	-	-	3	-	-	-	1	-
	Mod. Fam.	24h	6	-	9	4	4	-	12	1	1	-	1	-	23	-	-	-	1	-	-	-	1	-
	New Girl	24h	4	-	5	1	4	-	6	-	2	-	-	-	29	-	-	-	1	-	-	-	-	-
	Simpsons	24h	13	-	-	1	3	-	2	1	-	-	-	-	14	-	-	-	1	-	-	-	-	-
		Tot.	**22**	**5**	**9**	**2**	**4**	**4**	**12**	**62**	**2**	**0**	**4**	**2**	**44**	**0**	**1**	**0**	**4**	**0**	**0**	**0**	**1**	**0**
		ph	**0.229**	**0.052**	**0.075**	**0.021**	**0.042**	**0.042**	**0.125**	**0.517**	**0.021**	**0.0**	**0.042**	**0.021**	**0.458**	**0.0**	**0.010**	**0.0**	**0.042**	**0.0**	**0.0**	**0.0**	**0.010**	**0.0**
News	ABC	24h	-	-	3	-	-	-	2	9	1	-	1	-	11	-	-	-	1	-	-	-	-	-
	CBS	24h	12	1	1	1	-	-	7	24	-	-	-	-	13	-	-	-	1	-	-	-	1	-
	NBC	24h	2	4	-	-	2	1	-	23	1	-	2	2	6	-	-	-	2	-	-	-	-	-
	PBS	24h	8	-	5	1	2	3	3	6	-	-	1	-	14	-	1	-	-	-	-	-	-	-
		Tot.	**46**	**1**	**37**	**32**	**21**	**3**	**7**	**1**	**11**	**0**	**2**	**0**	**59**	**0**	**2**	**0**	**3**	**0**	**0**	**0**	**1**	**0**
		ph	**0.479**	**0.010**	**0.385**	**0.333**	**0.219**	**0.031**	**0.073**	**0.010**	**0.115**	**0.0**	**0.021**	**0.0**	**0.615**	**0.0**	**0.021**	**0.0**	**0.031**	**0.0**	**0.0**	**0.0**	**0.010**	**0.0**
Pro.	LibriSp.	24h	14	-	9	-	6	2	5	-	-	-	-	-	17	-	-	-	-	-	-	-	-	-
	Moz. CV	24h	10	1	21	5	14	1	2	1	11	-	2	-	18	-	1	-	1	-	-	-	-	-
	WSJ	24h	22	-	7	27	1	-	-	-	-	-	-	-	24	-	1	-	2	-	-	-	1	-
	CHiME	24h	1	-	3	3	-	-	10	2	7	-	1	-	1	-	-	-	1	-	-	-	-	-
Sum		Tot.	**100**	**6**	**80**	**43**	**43**	**9**	**67**	**67**	**23**	**0**	**9**	**2**	**198**	**0**	**4**	**0**	**15**	**0**	**0**	**0**	**5**	**0**
		ph	**0.321**	**0.019**	**0.256**	**0.138**	**0.138**	**0.029**	**0.215**	**0.215**	**0.074**	**0.0**	**0.029**	**0.006**	**0.635**	**0.0**	**0.0128**	**0.0**	**0.048**	**0.0**	**0.0**	**0.0**	**0.016**	**0.0**

A: Accidental triggers; *W*: Wake word said; Gray cells: Mismatch between played audio and wake word model language.

previous work [21], we added a row that describes the triggers *per hour* (ph) for each category.

In general, we cannot observe any noteworthy differences in accidental triggers (A) across the three dataset categories. In contrast, if we have a look at the cases where the wake word was actually said (W), we see that this was very often the case for "Computer" in the professional Wall Street Journal dataset caused by an article about the computer hardware company IBM and for "Amazon" across the news datasets. In this case, the 62 instances of "Amazon" referred 13 times to the 2019 Amazon rainforest wildfires and 49 times to the company.

If we look at the professional datasets, the number of triggers is within the same range or even increases compared to TV shows and news. As such, we have not found a speaker that triggered less often, because it might have been specifically trained on one of the professional datasets. In contrast to the other professional audio datasets, the CHiME dataset consists of recordings of group dinner scenarios resulting in comparatively less spoken words, explaining the overall lower number of accidental activations. Not presented in Table 3 is the *MUSAN* noise dataset, because we have not observed any triggers across the different speakers. This suggests that accidental triggers are less likely to occur for non-speech audio signals.

Comparison Between Local and Cloud-Based Triggers To reduce false device activations and improve wake word accuracy, smart speaker vendors make use of cloud-based verification systems (cf. Section 2.4.1). An overview of the distribution can be seen in Figure 4. Depending on the wake word, we find that the cloud ASR engine also misrecognizes about half of our accidental triggers. Fortunately for Cortana, only a small number of triggers (8 out of 197) are able to trick Microsoft's cloud verification.

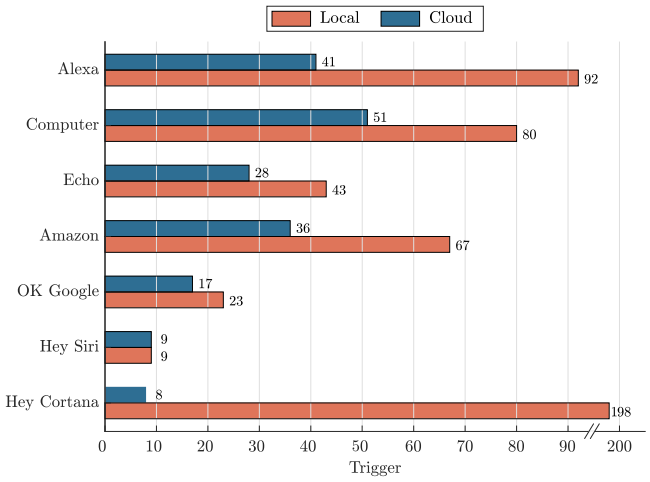

Figure 4: The number of accidental triggers that are incorrectly recognized by the local and the cloud-based ASR engine. Local triggers are triggers that are recognized as the wake word by the local model only. Cloud triggers are recognized by both the local and the cloud model.

Comparison Between Female and Male Speakers We performed an experiment designed to study a potential model bias in smart speakers, a common problem for machine learning systems [20, 40, 67]. We conducted a hypothesis test with $\alpha = .05$ using a Kruskal-Wallis H test, as the number of accidental triggers (discrete data) are not normally distributed (Shapiro-Wilk normality test, W = 0.66659, $p < 0.001$). Across our tested datasets, our experiments did not show a significant difference between female and male speakers (Kruskal-Wallis chi-squared = 0.16128, df = 1, p = 0.688). The detailed numbers are shown in Table 4.

Comparison Across Languages In Table 5, we report the results for the differences across languages to explore another potential model bias of the evaluated systems. Even though we only tested a small number of datasets per language, the number of triggers of VA5 Google and VA6 Apple is very low and comparable to their English performance. Given the fact that we played the very same episodes of the TV show *Modern Family* in English (US) and German, we find the wake word "Computer" to be more resistant to accidental triggers in German (1) than in English (9). A similar but less pronounced behavior can be seen with "Alexa." Moreover, we found that "big brother" in Standard Chinese dàgē (大哥) is often confused with the wake word "Echo", which is hence not the best wake word choice for this language. Similarly, the German words "Am Sonntag" ("On Sunday"), with a high prevalence notably in weather forecasts, are likely to be confused with "Amazon."

Multilingual Speakers and Language Mismatch Every year, the United States Census Bureau [68] surveys 3.5 million households across the US and publishes their findings in the American Community Survey (ACS). According to the ACS, 21.6% of the surveyed people speak a language other than English at home. With 13.4%, Spanish is the most common spoken language other than English in the US. Standard Chinese and German are spoken in less than $< 1\%$ of the surveyed US households. In our experiments, we found that the German and the three Chinese wake word models do not trigger very often on English (US) content (cf. right part of Table 3). However, more experiments with different languages are required to fully understand the privacy impact of accidental triggers on bilingual households.

4.3 Reproducibility

During the verification step of our accidental trigger search, we replayed every trigger 10 times to measure its reproducibility. This experiment is designed based on the insight that accidental triggers likely represent samples near the decision thresholds of the machine learning model. Furthermore, we cannot control all potential parameters during the empirical experiments, and thus we want to study if, and to which extent, a trigger is actually repeatable.

We binned the triggers into three categories: *low*, *medium*, and *high*. Audio snippets that triggered the respective assistant 1–3 times are considered as *low*, 4–7 times as *medium*, and 8–10 times as *high*. In Figure 5, we visualize these results.

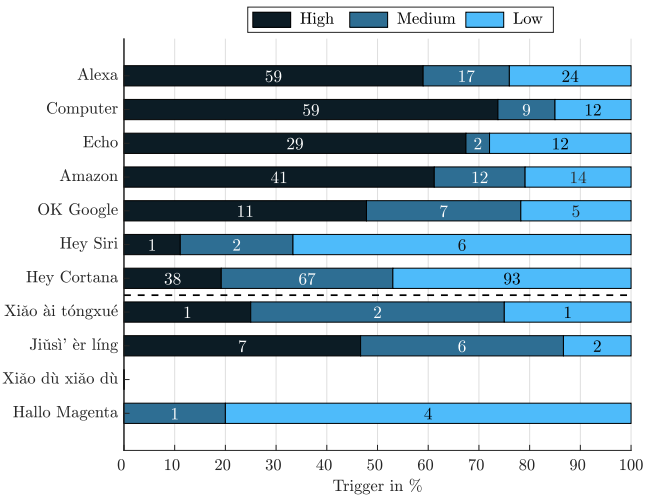

Figure 5: Accidental Trigger Reproducibility. Note that the four speakers below the dashed line do not use wake words in English (US); "Xiāo dù xiāo dù" did not have any triggers.

We observe that across the Amazon and Google speakers, around 75 % of our found triggers are medium to highly reproducible. This indicates that most of the identified triggers are indeed reliable and represent examples where the wake word recognition fails. For the Apple and Microsoft speakers, the triggers are less reliable in our experiments. One caveat of the results is that the Chinese and German speakers' data are rather sparse and do not allow any meaningful observation and interpretation of the results.

5 Crafting Accidental Triggers

The previous experiments raise the question of whether it is possible to specifically forge accidental triggers in a systematic and fully automated way. We hypothesize that words with a similar pronunciation as the wake word, i. e., based on similar phones (the linguistically smallest unit of sounds) are promising candidates. In this section, we are interested in crafting accidental triggers that are likely caused by the wake word's phonetic similarity.

5.1 Speech Synthesis

To systematically test candidates, we utilize Google's TTS API. To provide a variety across different voices and genders, we synthesize 10 different TTS versions, one for each US

Table 4: Differences Between Female and Male Speakers.

			Alexa		Computer		Echo		Amazon		Ok Google		Hey Siri		Hey Cortana		Xiǎo ài tóngxué		Jiǔsì' èr líng		Xiǎo dù xiǎo dù		Hallo Magenta	
			A	W	A	W	A	W	A	W	A	W	A	W	A	W	A	W	A	W	A	W	A	W
			en_us		en_us		en_us		en_us		en_us		en_us		en_us		zh_cn		zh_cn		zh_cn		de_de	
	Tot.		**31**	**3**	**9**	**0**	**4**	**6**	**10**	**0**	**0**	**0**	**0**	**0**	**41**	**0**	**1**	**0**	**1**	**0**	**0**	**0**	**2**	**0**
	ph		**0.646**	**0.062**	**0.188**	**0.0**	**0.083**	**0.125**	**0.208**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.875**	**0.0**	**0.020**	**0.0**	**0.020**	**0.0**	**0.0**	**0.0**	**0.042**	**0.0**
Fem.	LibriSp. I	24h	9	2	8	-	2	4	4	-	-	-	-	-	19	-	-	-	1	-	-	-	1	-
	LibriSp. II	24h	22	1	1	-	2	2	6	-	-	-	-	-	22	-	1	-	-	-	-	-	1	-
	Tot.		**33**	**0**	**8**	**0**	**0**	**8**	**8**	**2**	**1**	**0**	**0**	**0**	**46**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	ph		**0.688**	**0.0**	**0.167**	**0.0**	**0.0**	**0.167**	**0.167**	**0.042**	**0.021**	**0.0**	**0.0**	**0.0**	**0.958**	**0.0**	**0.021**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Male	LibriSp. I	24h	19	-	3	-	-	4	5	2	-	-	-	-	20	-	1	-	-	-	-	-	-	-
	LibriSp. II	24h	14	-	5	-	-	4	3	-	1	-	-	-	26	-	-	-	-	-	-	-	-	-

A: Accidental triggers; *W*: Wake word said; Gray cells: Mismatch between played audio and wake word model language.

Table 5: Differences in Languages.

		Alexa		Computer		Echo		Amazon		Ok Google		Hey Siri		Hey Cortana		Xiǎo ài tóngxué		Jiǔsì' èr líng		Xiǎo dù xiǎo dù		Hallo Magenta	
		A	W	A	W	A	W	A	W	A	W	A	W	A	W	A	W	A	W	A	W	A	W
English	**Time**	**en_us**		**en_us**		**en_us**		**en_us**		**en_us**		**en_us**		**en_us**		zh_cn		zh_cn		zh_cn		de_de	
Modern Family	24h	6	-	9	4	4	-	12	1	1	-	1	-	23	-	-	-	1	-	-	-	1	-
German	**Time**	**de_de**		**de_de**		**de_de**		**de_de**		**de_de**		**de_de**		en_us		zh_cn		zh_cn		zh_cn		**de_de**	
Modern Family	24h	1	1	1	13	3	-	13	1	2	-	2	-	17	-	-	-	1	-	-	-	-	-
Big Bang Theory	24h	-	-	1	9	9	-	3	2	2	1	1	1	12	-	-	-	-	-	-	-	1	-
Polizeiruf 110	24h	3	-	4	7	3	-	13	-	-	-	-	-	18	-	-	-	-	-	-	-	-	-
Tatort	24h	-	-	-	8	4	-	15	1	2	-	-	-	6	-	-	-	-	-	-	-	-	-
ARD Tagesschau	12h	3	-	1	1	-	-	10	13	1	-	-	1	29	-	1	-	-	-	-	-	-	-
ZDF Heute Journal	12h	-	-	-	4	-	-	5	3	-	-	-	-	8	-	-	-	1	-	-	-	-	-
Standard Chinese	**Time**	en_us		en_us		en_us		en_us		en_us		en_us		en_us		**zh_cn**		**zh_cn**		**zh_cn**		de_de	
All Is Well	24h	1	-	1	-	9	-	6	-	-	-	-	-	28	-	-	-	-	-	2	-	-	-
CCTV X. Lianbo	12h	3	-	1	-	1	-	-	-	-	-	1	-	38	-	-	-	3	-	-	-	-	-

A: Accidental triggers; *W*: Wake word said; Gray cells: Mismatch between played audio and wake word model language.

English voice in the TTS API. Four of the voices are standard TTS voices; six are Google *WaveNet* voices [70]. In both cases, the female-male-split is half and half.

Note that some words have more than one possible pronunciation (e. g., T AH M **EY** T OW vs. T AH M **AA** T OW). Unfortunately, we cannot control how Google's TTS service pronounces these words. Nevertheless, we are able to show how, in principle, one can find accidental triggers, and we use 10 different voices for the synthesis to limit this effect.

5.2 Levenshtein Distance

To compare the wake words with other words, we use the Fisher corpus [15] version of the Carnegie Mellon University pronouncing dictionary [44], an open-source pronunciation dictionary for North American English listing the phone sequences of more than 130,000 words. We propose two versions of a weighted phone-based Levenshtein distance [50] to measure the distance of the phonetic description of a candidate to the phonetic description of the respective wake word in order to find potential triggers in a fully automated way. Using dynamic programming, we can compute the minimal distance $\mathcal{L}$ (under an optimal alignment of the wake word and the trigger word). Formally, we calculate

$$\mathcal{L} = \frac{s \cdot S + d \cdot D + i \cdot I}{N} \tag{1}$$

with the number of *substituted* phones S, *inserted* phones I, *deleted* phones D, and the total number of phones N, describing the weighted *edit distance* to transform one word into another. The parameters s, d, and i describe scale factors for the different kinds of errors.

In the following, we motivate our different scale factors: During the decoding step of the recognition pipeline, a path

search through all possible phone combinations is conducted by the automatic speech recognition system. In general, for the recognition, the path with the least cost is selected as the designated output of the recognition (i. e., wake word or not wake word). Considering these principles of wake word recognition, we assume that the different kinds of errors have different impacts on the wake word recognition, as e. g., utterances with deletions of relevant phones will hardly act as a wake word.

To find the optimal scale factors, we conducted a *hyperparameter search* where we tested different combinations of weights. For this, we played all different TTS versions of 50,000 English words and measured which of the voice assistants triggered at least once. In total, we were able to measure 826 triggers. In a second, more advanced, version of this distance measure, we considered phone-depended weights for the different kinds of errors. A more detailed description of this version of the distance measure is presented in Section 5.3.

We ignore words which are either the wake word itself or pronounced like parts of the wake word (e. g., "Hay" is blocklisted for "Hey" or "computed" for "computer"). The blocklist of the wake words contains a minimum of 2 words (*Cortana*) and up to 6 words (*Computer*). For the optimization, we used a ranked-based assessment: We sorted all $50,000$ words by their distance $\mathcal{L}$ and used the rank of the triggered word with the largest distance as a metric to compare the different weighted Levenshtein distances. With this metric, we performed a grid search for s, d, i over the interval $[0, 1]$ with a step width of 0.05.

Note that not all accidental triggers can be explained effectively with the proposed model. Therefore, in a first step, we filter all available triggers to only include those that can be described with this model. This step is necessary, as we are not interested in crafting *all* possible accidental triggers such as noise, but accidental triggers that are likely caused by the phonetic similarity to the wake word only. Also, the Invoke speaker and the Google Home Mini both have two potential wake words. By focusing on the subset of accidental triggers that can describe the respective wake word more closely, we can filter out the other version of the wake word. Specifically, we only used triggers were we were able to describe the trigger with the proposed distance measure in such a way that it remained within the first 1 % (500) of words if we overfitted the distance measure to that specific word. In other words, we only considered triggers for our hyperparameter search where a combination of scale factors exits such that the trigger has at most the rank 500. After applying this filter criterion, 255 out of the 826 triggers remained in the dataset.

5.3 Phone-Dependent Weights

For a more advanced version of the weighted Levenshtein distance, we utilized information about how costly it is to substitute, delete, and insert specific phones (i. e., intuitively it should be less costly to replace one vowel with another vowel in comparison to replacing a vowel with a consonant). For this, we calculated phone-dependent weights as described in the following: We used a trained ASR system and employed *forced alignment*, which is usually used during the training of an ASR system to avoid the need for a detailed alignment of the transcription to the audio file. We can use this algorithm to systematically change phones in the transcription of an audio file and measure the costs of these specific changes.

To measure the impact of such changes, we distinguish between deletions, substitutions, and insertions: To assess the cost of the deletion of specific phones, we randomly draw 100 words that contain that specific phone and synthesize 10 versions of this word via Google's TTS API. We use the difference of the scores of the forced alignment output with and without this specific phone for all TTS versions of the word. For example, we use the word *little* with the phonetic description L IH T AH L for the phone AH in *Alexa* and measure the score of the forced alignment algorithm for L IH T AH L and L IH T L. The loss in these two scores describes the cost of deleting the sound 'AH' in this specific context. For the final weights, we use the average over all 100 words and 10 TTS versions and finally normalize the values of all averaged phone costs to obtain a mean value of 1.0. The resulting deletion weights $\hat{d}$ are shown in Figure 6.

Similarly, to determine the cost of all possible substitutions, we replace the phone-under-test with all other phones for all 100 words and 10 TTS versions. We followed the same approach as for the deletion costs, averaging and normalizing the log-likelihood scores to define the final weights. The matrix of the substitution weights is shown in Figure 7. The rows describe the original phones of the wake words and the columns the substituted phone. The higher the value, the higher are the costs for the transcription if the phone of a wake word is replaced by the respective other phone. Note that we only calculated the weights of phones that occur in the wake words. Therefore, the rows in the figure do not show all possible phones of the language. The rows of the matrix are also normalized to have an average value of 1.0. Finally, we compare the scores between the original transcription and insert the considered phone for the insertion weights. These weights are also normalized to have an average value of 1.0. The insertion weights are shown in Figure 8. All weights are then used along with the scale factors.

5.4 Cross-Validation

We performed a leave-one-out cross-validation to measure the performance of Equation (1) in predicting whether words are potential accidental triggers. For this purpose, we compared three different versions of Equation (1): a version, with all scales set to 1 (Unweighted), a scaled version where we optimized the scale factors (Simple), and a version where we

used our optimized scale factors and the phone-dependent weights (Advanced).

Table 6: Results of the leaving-one-out cross-validation. We report the numbers of triggers within the 100 words with the smallest distance to the respective wake word.

ID	Wake Word	Total	Unweighted	Simple	Advanced
VA1	*Alexa*	52	9	17	24
VA2	*Computer*	75	17	21	32
VA3	*Echo*	23	4	5	12
VA4	*Amazon*	12	1	1	7
VA5a	*OK Google*	2	0	0	0
VA5b	*Hey Google*	1	0	0	0
VA6	*Hey Siri*	7	3	3	5
VA7a	*Hey Cortana*	38	9	9	6
VA7b	*Cortana*	45	13	14	10

We have run a hyperparameter search for the simple and the advanced version of eight wake words triggers for each fold and tested the resulting scale factors on the remaining wake word. The results in Table 6 show the number of triggers we find within the 100 words with the smallest distance for all three versions of the Levenshtein distance and all wake words. Note that the distances tend to cluster words into same distances due to the fixed length of each wake word and, therefore, the same total number of phones N, especially for the unweighted and the simple version.

For cases where it is not possible to clearly determine the closest 100 words, we use all words with a smaller distance than the 100^{th} word and draw randomly out of the words with the next largest distance until we obtain a list of 100 words which makes sure to have a fair comparison in Table 6.

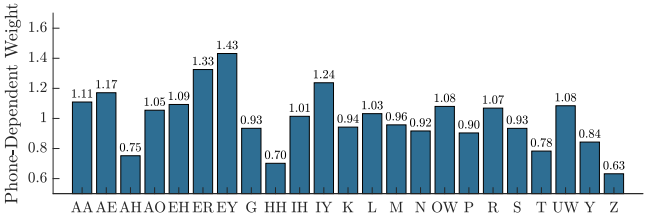

Figure 6: Deletion weights used for the advanced version of the weighted Levenshtein distance. The higher the value, the higher the costs if this phone is *removed*.

In the third column (Total), we show the total number of words that triggered the perspective wake words, out of the 50,000 words, after filtering. Note that the Google wake words had only 1 or 2 triggers and that, therefore, not more than these can be in the top 100.

The different versions of the Levenshtein distance generally show better results for the simple and the advanced version compared to the unweighted version, especially for all Amazon wake words. Only for the two wake words from Microsoft, this is not the case. Nevertheless, the advanced

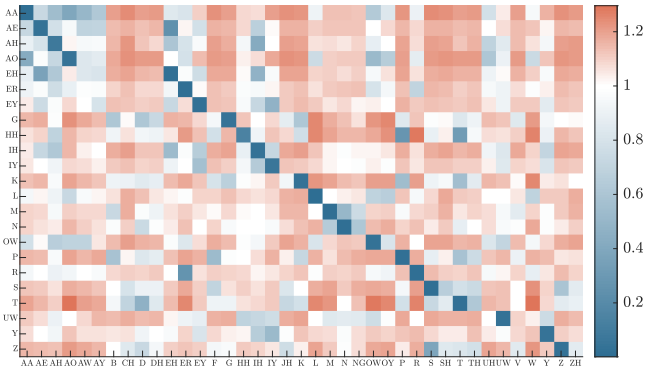

Figure 7: Substitution weights used for the advanced version of the weighted Levenshtein distance plotted as a matrix describing the cost to *replace* the phone in the row with a phone of the columns.

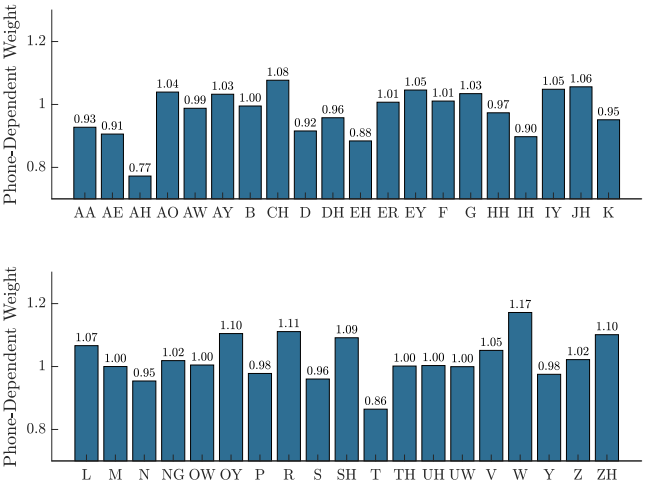

Figure 8: Insertion weights used for the advanced version of the weighted Levenshtein distance. The higher the value, the higher the costs if this phone is *inserted*.

version shows the best results on average and is, therefore, the version we use in the following experiments. Notably, for e. g., *Computer*, approximately one third (32/100) of the words with the smallest distance actually triggered the smart speaker and for many of the wake words, more or almost half of all possible triggers can be found within the 100 words with the smallest distance.

5.5 Performance on Real-World Data

With the optimized scale factors and weights, we evaluate the distance measure on the transcriptions of the CHiME dataset to assess the performance of the optimized distance measure on real-world voice data. For this purpose, we consider n-grams to test also sequences of words that occur in the CHiME transcriptions, namely 1-, 2-, and 3-grams.

We perform a hyperparameter search for the advanced version of the Levenshtein distance (scale factors and phone-dependent weights) on the triggers of all 9 wake words on the data set used in Section 5.4. For these, the optimal scale factors are $s = 1.46$, $d = 1.30$, and $i = 0.24$, which we use in the following experiment. We select the 100 words with the smallest distance to the respective wake word from all 1-, 2-, and 3-grams. In total, 300 n-grams for each wake word. All these n-grams are synthesized with Google's TTS API. We then play these crafted triggers against all smart speakers. The results of the CHiME n-grams are shown in Table 7.

We have identified very common words like "compare" (*Computer*; rank 405), "collection" (*Alexa*; rank 743), "technical" (*Echo*; rank 775), or "ago" (*Echo*; rank 807) and phrases like "collection of" for *Alexa*, "compared to" for *Computer*, "capable of" for *OK Google*, "a goal" for *Echo*, "fresh parmesan" for *Amazon*, "my cereal" for *Hey Siri*, and "all acts of" for *Alexa*. A manual analysis revealed that some of the crafted triggers were also found, and also caused triggers, in our previously tested real-world data. Examples include: (Alexa) "*Election* day is" in Modern Family, (Computer) "*compared to* the millions" in PBS NewsHour, and (Cortana) "*Montana*'s bag limits are" in LibriSpeech.

Table 7: We construct word sequences based on n-grams from the CHiME transcriptions. We report the numbers of triggers within the 100 n-grams with the smallest distance to the respective wake word.

ID	Wake Word	1-gram	2-gram	3-gram
VA1	*Alexa*	7	10	5
VA2	*Computer*	16	12	10
VA3	*Echo*	1	8	3
VA4	*Amazon*	2	11	4
VA5a	*OK Google*	0	1	0
VA5b	*Hey Google*	0	0	0
VA6	*Hey Siri*	2	2	0
VA7a	*Hey Cortana*	8	8	4
VA7b	*Cortana*	7	5	6

6 Related Work

There is an increasing amount of work focusing on the privacy of smart speakers that motivates and guides our research, as discussed in the following.

6.1 Smart Speaker Privacy

Malkin et al. [46] studied the privacy attitudes of 116 smart speaker users. Almost half of their respondents did not know that their voice recordings are stored in the cloud, and only a few had ever deleted any of their recordings. They reported that their participants were particularly protective about other people's recordings, such as guests. Besides conversations that include children, financial, sexual, or medical information, *accidentally captured conversations* were named information that should automatically be screened out and not stored. Lau et al. [42] studied privacy perceptions and concerns around smart speakers. They found an incomplete understanding of the resulting privacy risks and document problems with incidental smart speaker users. For example, they describe that two of their participants used the audio logs to surveil or monitor incidental users. They noted that current privacy controls are rarely used. For example, they studied why users do not make use of the mute button on the smart speaker. Most of their participants preferred to simply unplug the device and give trust issues and the inability to use the speaker hands-free as reasons not to press the mute button.

Similarly, Abdi et al. [1] explored mental models of where smart speaker data is stored, processed, and shared. Ammari et al. [5] studied how people use their voice assistants and found users being concerned about *random activations* and documented how they deal with them. Huang et al. [35] studied users' concerns about shared smart speakers. Their participants expressed worries regarding *voice match false positives*, unauthorized access of personal information, and the misuse of the device by unintended users such as visitors. They confirmed that users perceive external entities, such as speaker vendors, collecting voice recordings as a major privacy threat. Chung et al. [14] named *unintentional voice recordings* a significant privacy threat and warned about entities with legitimate voice data access and commercial interests, as well as helpless users not in control of their voice data. Tabassum et al. [66] studied *always-listening* voice assistants that do not require any wake word. Zeng et al. [74] studied security and privacy-related attitudes of people living in smart homes. Their participants mentioned privacy violations and concerns, particularly around audio recordings. In this paper, we study the actual prevalence and implications of accidental triggers with the goal of providing tangible data on this phenomenon, as well as an effective process for assessing trigger accuracy of devices by means of crafting likely accidental triggers.

Dubois et al. [21] published a paper where they played 134 hours of TV shows to measure the prevalence of accidental triggers. Their setup relied on a combination of a webcam, computer vision, and a network traffic-based heuristic. In contrast to our work, the authors focused only on a comparatively small TV show dataset and English-speaking smart speakers. They did not consider speakers from other countries, other languages, or other audio datasets. Furthermore, while their work only speculates about regional differences, our reverse engineering of Amazon Echo internals confirms the existence of different wake word models per language, region, and device type (e. g., en-US vs. en-GB). Finally, we propose a method to craft accidental triggers that enables us to find new triggers systematically and discuss possible countermeasures.

6.2 Inaudible and Adversarial Examples

Adversarial examples against speech recognition systems try to fool the system to output a wrong transcription. For human listeners, the adversarial examples are not at all or only barely distinguishable from benign audio. In 2016, Carlini et al. [13] have shown that targeted attacks against HMM-only ASR systems are possible. To create their adversarial audio samples, they used an inverse feature extraction. The resulting audio samples were not intelligible by humans. Schönherr et al. [59] presented an approach where psychoacoustic modeling, which is borrowed from the MP3 compression algorithm, was used to re-shape the perturbations of the adversarial examples. Their approach improves previous work by hiding the changes to the original audio below the human hearing thresholds. Later, the attack was ported to an over-the-air setting by crafting examples that remain robust across different rooms [58]. The accidental triggers identified by our work can be combined with adversarial examples to wake up smart speakers in an inconspicuous way.

7 Discussion

In this section, we discuss and interpret our findings and propose countermeasures that can help to reduce the impact of accidental triggers.

7.1 Prevalence and Privacy Impact

Overall, the number of observed accidental triggers differ across smart speaker vendor and wake word model. For the popular English US Alexa wake word model and the most realistic of the tested scenarios, i. e., TV shows, we observed one accidental trigger every 4 hours, so, depending on the usage, possibly multiple times a day. Whether this is reason for concern or perceived as low and acceptable ultimately relies on the user. However, from the literature, it is known that misactivations are perceived as "a major privacy threat" for some users and may also affect unintended users such as visitors [5, 14, 21, 35, 46]. Moreover, it is documented that users have an incomplete understanding of the privacy risks and that existing privacy controls are rarely used [42]. As the underlying problem of accidental triggers, the trade-off between a low false acceptance and false rejection rate is hard to balance, we will discuss potential measures that can help to reduce the impact of accidental triggers on users' privacy in the following.

7.2 Wake Word

The results of our experiments suggest possible reasons for the differences across smart speakers and raise the question about the importance of the wake word and why their vendors have chosen them in the first place.

Properties of Robust Wake Words Looking at the number of words in a wake word, one would assume a clear benefit using two words. This observation is supported by the results in Table 6, where "Cortana" leads to more triggers than "Hey Cortana." On the contrary, the shortest wake word "Echo" has fewer triggers than "Hey Cortana," suggesting that not only the number of words (and phones) itself is important, but the average distance to common words in the respective language. These results suggest that increasing the number of words in a wake word has the same effect as increasing the distance to common words. If we consider the differences in the prevalence of accidental triggers, and that adding an additional word (e. g., "Hey") comes at close to no cost for the user, we recommend that vendors deploy wake words consisting of two words.

Word Selection Amazon shared some details about why they have chosen "Alexa" as their wake word [12]: The development was inspired by the LCARS, the Star Trek computer, which is activated by saying "Computer." Moreover, they wanted a word that people do not ordinarily use in everyday life. In the end, Amazon decided on "Alexa" because it sounded unique and used soft vowels and an "x." The cofounder of Apple's voice assistant chose the name "Siri" after a co-worker in Norway [32]. Later, when Apple turned Siri from a push-to-talk into a wake word-based voice assistant, the phrase "Hey Siri" was chosen because they wanted the wake word to sound as natural as possible [7]. Based on those examples we can see that the wake word choice in practice is not always a rational, technically founded decision, but driven by other factors like marketing as in "OK Google," "Amazon," "Xiǎo dù xiǎo dù," or "Hallo Magenta," or based on other motivations such in the case of "Siri" or "Computer." Another issue can arise when trying to port a wake word across languages. An example of that is the confusion of dàgē ("big brother") and "Echo" described in Section 4.2, and it gets even more complicated in multilingual households [64].

7.3 Countermeasures

As long as the precise detection of wake words remains a challenge, there is a need for preventing and limiting the impact of accidental triggers.

Limiting the Impact As stated in their privacy policy, the Magenta Speaker from Deutsche Telekom automatically stops the upload of the audio stream after 8 seconds if no voice command can be detected in the recording [19].

While such accidental uploads are still problematic, it is an easy-to-implement protection mechanism, which has the potential to at least limit the privacy violation caused by accidental triggers. Similarly, Google's feature that allows users to adjust the wake word's responsiveness might help in certain environments (cf. Section 2.2).

Local On-Device Speech Recognition Coucke et al. [17] describe a smart speaker that runs completely offline and is thus private-by-design. In 2019, Google deployed an on-device speech recognizer on a smartphone that can transcribe spoken audio in real-time without an Internet connection [31, 37]. We find such an approach to be promising, as it can help to reduce the impact of accidental triggers by limiting the upload of sensitive voice data. After the local ASR detects the wake word, one can imagine a speaker that transcribes the audio and only after being ensured to have detected a user command, uploads the short wake word sequence for cloud verification. When both ASR engines agree about the wake word's presence, the command is forwarded to the cloud in text or audio form. Ahmed et al. [3] describe a speech transcription service that applies a series of privacy-protecting operations before uploading the voice data to the cloud, but in its current form, also introduces a very high latency. Sigtia et al. [61] explore the accuracy vs. latency tradeoff that exists when including parts of the audio following the wake word as a signal to detect accidental triggers.

Device-Directed Queries and Visual Cues Amazon presented a classifier for distinguishing device-directed queries from background speech in the context of follow-up queries [4, 47]. While follow-up queries are a convenience feature, one can imagine a similar system that can reduce the number of accidental triggers. Mhaidli et al. [48] explored the feasibility to only selectively activate a voice assistant using gaze direction and voice volume level by integrating a depth-camera to recognize a user's head orientation. While this approach constitutes a slight change in how users interact with a smart speaker, it effectively reduces the risk of accidental triggers, by requiring a direct line-of-sight between the user and the device. However, their participants also expressed privacy concerns due to the presence of the camera.

Privacy Mode and Safewords Lau et al. [42] has documented the ineffectiveness of current privacy controls, such as the mute button, given the inability to use the speaker hands-free when muted. We imagine a method similar to a *safeword* as a possible workaround for this problem. For this, the speaker implements a *privacy mode* that is activated by a user saying, "Alexa, please start ignoring me," but could, for example, also be activated based on other events such as the time of the day. In the privacy mode, the speaker disables all cloud functionality, including cloud-based wake word verification and question answering.

The speaker's normal operation is then re-enabled by a user saying, "Alexa, Alexa, Alexa." Repeating the wake word multiple times is similar to a behavior observed when parents call their children multiple times, if they do not like to listen, this will feel natural to use. Due to the requirement to speak the somewhat lengthy safeword, accidental triggers will only happen very rarely. We imagine this privacy control to be more usable than a mute button, as the hands-free operation is still possible. As only the wake word is repeated multiple times, we think that vendors can implement this functionality using the local ASR engine.

Increased Transparency & Informed Consent Another option is to increase transparency and control over the retention periods and individual uploads and recordings. In particular, our experience with Microsoft's *Privacy Dashboard* made it clear that vendors need to implement features to better control, sort, filter, and delete voice recordings. Amazon's and Google's web interface already allow a user to filter interactions by date or device easily. In particular, we imagine a view that shows potential accidental triggers, e. g., because the assistant could not detect a question. Currently, accidental triggers are (intentional) not very present, and are easy to miss in the majority of legitimate requests. If accidental triggers are more visible, we hope that users will start to more frequently use privacy controls such as safewords, the mute button, or to request the deletion of the last interaction via a voice command, e. g., "Hey Google, that wasn't for you." At first, integrating such a functionality seems unfavorable to vendors, but it can easily be turned into a privacy feature that can be seen as an advantage over competitors.

Similarly, how the voice data collection practices are communicated with the user and the ways of obtaining *informed* consent need to be improved. While Google turned off the voice data collection for every user in August 2020 [11] by default, the company also added a clarifying video that better explains why and how the voice data is collected and manually reviewed by third parties. In contrast, Amazon still "hides" their practices in the terms of service, and Apple nudges users to give their consent via periodic reminders and the specific design of the consent form in their user interface.

Discriminative Training with Crafted Wake Words Crafting accidental triggers may also be used to reinforce wake word detection, for example, by discriminative training. Here, the likelihood ratio (or more generally, the system-specific discriminance score) between the actual wake words and all potentially confusing word sequences should be maximized. These potential confounders could be identified by crafting triggers as described in Section 5.

7.4 Limitations

We have neither evaluated nor explored triggers for varying rooms and acoustic environments, e. g., distances or volumes. Varying conditions influence the propagation in a room and, therefore, the recognition of the audio signal. Even if this might influence the reproducibility, this was not part of our study, as we focused on the general number of accidental triggers in a comparable setup across all experiments. This also implies that our results are somewhat tied to the hard-

and software version of the evaluated smart speakers. Our results are subject to change due to model updates for the local ASR or updates of the cloud model. Furthermore, we are dealing with a system that is not entirely deterministic, as others already noted [41]. Accidental triggers we mark as local triggers sometimes pass the cloud-based recognizer and vice versa. Our findings are mainly based on the English (US) language; even though we also played a limited set of German and Standard Chinese media, our results are not generalizable to other languages or ASR models. This is mainly because spoken languages can be different by accent, vocal pitch, tone, word stress, and more. For example, Standard Chinese is a tonal language, where the meaning of a syllable is affected by its pitch trajectory. German is more closely related to English, but makes more use of compound words and differs in the number and identity of its phonetic units.

8 Conclusion

In this work, we conduct a comprehensive analysis of accidental triggers in voice assistants and explore their impact on the user's privacy. We explain how current smart speakers try to limit the impact of accidental triggers using cloud-based verification systems and analyze how these systems affect users' privacy. More specifically, we automate the process of finding accidental triggers and measure their prevalence across 11 smart speakers. We describe a method to artificially craft such triggers using a pronouncing dictionary and a weighted phone-based Levenshtein distance metric that can be used to benchmark smart speakers. As the underlying problem of accidental triggers, the trade-off between a low false acceptance and false rejection rate is hard to balance. We discuss countermeasures that can help to reduce the number and impact of accidental triggers. To foster future research on this topic, we publish a data set of more than 350 accidental triggers.

Acknowledgments

We thank Wentao Yu and Miranda Wei for their assistance and help with Standard Chinese translations. This research was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2092 *CASA* – 390781972.

Availability

We publish a dataset of more than 350 accidental triggers to foster future research on this topic. They are available at https://unacceptable-privacy.github.io, where we also provide example videos.

References

[1] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. More than Smart Speakers: Security and Privacy Perceptions of Smart Home Personal Assistants. In *Symposium on Usable Privacy and Security*, SOUPS '19, pages 451–466, Santa Clara, California, USA, August 2019. USENIX.

[2] Saurabh Adya, Vineet Garg, Siddharth Sigtia, Pramod Simha, and Chandra Dhir. Hybrid Transformer/CTC Networks for Hardware Efficient Voice Triggering. In *Conference of the International Speech Communication Association*, INTERSPEECH '20, pages 3351–3355, Virtual Conference, October 2020. ISCA.

[3] Shimaa Ahmed, Amrita Roy Chowdhury, Kassem Fawaz, and Parmesh Ramanathan. Preech: A System for Privacy-Preserving Speech Transcription. In *USENIX Security Symposium*, SSYM '20, pages 2703–2720, Virtual Conference, August 2020. USENIX.

[4] Amazon, Inc. Alexa: Turn on Follow-Up Mode, February 2020. https://www.amazon.com/gp/help/customer/display.html?nodeId=202201630, as of November 15, 2021.

[5] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Transactions on Computer-Human Interaction*, 26(3):17:1–17:28, June 2019.

[6] Apple, Inc. Hey Siri: An On-device DNN-powered Voice Trigger for Apple's Personal Assistant, October 2017. https://machinelearning.apple.com/2017/10/01/hey-siri.html, as of November 15, 2021.

[7] Apple, Inc. Personalized Hey Siri, April 2018. https://machinelearning.apple.com/2018/04/16/personalized-hey-siri.html, as of November 15, 2021.

[8] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines. In *Annual Conference of the International Speech Communication Association*, INTERSPEECH '18, pages 1561–1565, Hyderabad, India, September 2018. ISCA.

[9] Mark Barnes. Alexa, Are You Listening?, August 2017. https://labs.f-secure.com/archive/alexa-are-you-listening/, as of November 15, 2021.

[10] Dieter Bohn. Amazon Says 100 Million Alexa Devices Have Been Sold, January 2019. https://www.theverge.com/2019/1/4/18168565/, as of November 15, 2021.

[11] Dieter Bohn. Google Is Sending a Complicated Privacy Email to Everyone, August 2020. https://www.theverge.com/2020/8/5/21354805/, as of November 15, 2021.

[12] Julie Bort. Amazon Engineers Had One Good Reason and One Geeky Reason for Choosing the Name Alexa, July 2016. https://www.businessinsider.com/why-amazon-called-it-alexa-2016-7, as of November 15, 2021.

[13] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden Voice Commands. In *USENIX Security Symposium*, SSYM '16, pages 513–530, Austin, Texas, USA, August 2016. USENIX.

[14] Hyunji Chung, Michaela Iorga, Jeffrey Voas, and Sangjin Lee. "Alexa, Can I Trust You?". *IEEE Computer*, 50(9):100–104, September 2017.

[15] Christopher Cieri, David Miller, and Kevin Walker. The Fisher Corpus: A Resource for the Next Generations of Speech-to-Text. In *International Conference on Language Resources and Evaluation*, LREC '04, pages 69–71, Lisbon, Portugal, May 2004. ELRA.

[16] Ike Clinton, Lance Cook, and Shankar Banik. A Survey of Various Methods for Analyzing the Amazon Echo, August 2016. https://vanderpot.com/Clinton_Cook_Paper.pdf, as of November 15, 2021.

[17] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. Snips Voice Platform: An Embedded Spoken Language Understanding System for Private-by-Design Voice Interfaces. *CoRR*, abs/1805.10190:1–29, May 2018.

[18] Matt Day, Giles Turner, and Natalia Drozdiak. Amazon Workers Are Listening to What You Tell Alexa, April 2019. https://www.bloomberg.com/news/articles/2019-04-10/is-anyone-listening-to-you-on-alexa-a-global-team-reviews-audio, as of November 15, 2021.

[19] Deutsche Telekom, AG. Privacy Policy: Magenta Smart Speaker, July 2020. https://www.telekom.de/datenschutzhinweise/download/181.pdf, as of November 15, 2021.

[20] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and Mitigating Unintended Bias in Text Classification. In *AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 67–73, New Orleans, Louisiana, USA, February 2018. ACM.

[21] 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. In *Privacy Enhancing Technologies Symposium*, PETS '20, pages 255–276, Virtual Conference, July 2020. Sciendo.

[22] Mozilla Foundation and Community. Mozilla: Common Voice, June 2017. https://voice.mozilla.org, as of November 15, 2021.

[23] Chaim Gartenberg. Apple Apologizes for Siri Audio Recordings, Announces Privacy Changes Going Forward, August 2019. https://www.theverge.com/2019/8/28/20836760/, as of November 15, 2021.

[24] Andrew Gebhart. Is Google Home Good at Voice Recognition?, April 2017. https://www.cnet.com/news/is-google-home-good-at-voice-recognition/, as of November 15, 2021.

[25] Google, Inc. Google Assistant Sensitivity, April 2020. https://www.blog.google/products/assistant/more-ways-fine-tune-google-assistant-you/, as of November 15, 2021.

[26] Google, Inc. Google Assistant with Voice Match – Upgraded Voice Match, February 2020. https://support.google.com/assistant/answer/9071681, as of November 15, 2021.

[27] Google, Inc. Adjust How Sensitive Google Assistant Is to "Hey Google", July 2021. https://support.google.com/assistant/answer/9712065, as of November 15, 2021.

[28] Google, Inc. Get Sound After You Say "Hey Google", July 2021. https://support.google.com/googlenest/answer/7410241, as of November 15, 2021.

[29] Jinxi Guo, Kenichi Kumatani, Ming Sun, Minhua Wu, Anirudh Raju, Nikko Ström, and Arindam Mandal. Time-Delayed Bottleneck Highway Networks Using a DFT Feature for Keyword Spotting. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '18, pages 5489–5493, Calgary, Alberta, Canada, April 2018. IEEE.

[30] Jaap Haitsma and Ton Kalker. A Highly Robust Audio Fingerprinting System. In *International Conference on Music Information Retrieval*, ISMIR '02, pages 1–9, Paris, France, October 2002. IRCAM.

[31] Yanzhang He, Tara N. Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, and Ding Zhao *et al.* Streaming End-to-end Speech Recognition for Mobile Devices. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '19, pages 6381–6385, Brighton, United Kingdom, May 2019. IEEE.

[32] Yoni Heisler. Steve Jobs Wasn't a Fan of the Siri Name, March 2012. https://www.networkworld.com/article/2221246/, as of November 15, 2021.

[33] Alex Hern. Alexa Users Can Now Disable Human Review of Voice Recordings, August 2019. https://www.theguardian.com/technology/2019/aug/05/alexa-allows-users-to-disable-human-review-of-voice-recordings, as of November 15, 2021.

[34] Alex Hern. Apple Contractors 'Regularly Hear Confidential Details' on Siri Recordings, July 2019. https://www.theguardian.com/technology/2019/jul/26/apple-contractors-regularly-hear-confidential-details-on-siri-recordings, as of November 15, 2021.

[35] Yue Huang, Borke Obada-Obieh, and Konstantin Beznosov. Amazon vs. My Brother: How Users of Shared Smart Speakers Perceive and Cope with Privacy Risks. In *ACM Conference on Human Factors in Computing Systems*, CHI '20, pages 402:1–13, Honolulu, Hawaii, USA, April 2020. ACM.

[36] Ted Karczewski. Cloud-Based Wake Word Verification Improves "Alexa" Wake Word Accuracy, May 2017. https://developer.amazon.com/en-US/docs/alexa/alexa-voice-service/enable-cloud-based-wake-word-verification.html, as of November 15, 2021.

[37] Jacob Kastrenakes. Pixel 4 Recorder App Can Transcribe Speech in Real-Time Without an Internet Connection, October 2019. https://www.theverge.com/2019/10/15/20915452/, as of November 15, 2021.

[38] Bret Kinsella. Deutsche Telekom and SoundHound Make Their Partnership Public with Houndify Supporting Magenta, October 2019. https://voicebot.ai/2019/10/31/deutsche-telekom-and-soundhound-make-their-partnership-public-with-houndify-supporting-magenta/, as of November 15, 2021.

[39] Bret Kinsella. The Bedroom is Now the Most Popular Location for Smart Speakers, April 2020. https://voicebot.ai/2020/04/30/yes-the-bedroom-is-now-the-most-popular-location-for-smart-speakers-heres-why-and-what-it-means/, as of November 15, 2021.

[40] Svetlana Kiritchenko and Saif M. Mohammad. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Conference on Lexical and Computational Semantics*, *SEM '18, pages 43–53, New Orleans, Louisiana, USA, June 2018. ACL.

[41] Deepak Kumar, Riccardo Paccagnella, Paul Murley, Eric Hennenfent, Joshua Mason, Adam Bates, and Michael Bailey. Skill Squatting Attacks on Amazon Alexa. In *USENIX Security Symposium*, SSYM '19, pages 33–47, Santa Clara, California, USA, August 2019. USENIX.

[42] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-seeking Behaviors with Smart Speakers. In *ACM Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW '18, pages 102:1–102:31, New York City, New York, USA, November 2018. ACM.

[43] Colin Lecher. Google Will Pause Listening to EU Voice Recordings While Regulators Investigate, August 2019. https://www.theverge.com/2019/8/1/20750327/, as of November 15, 2021.

[44] Kevin A. Lenzo. Carnegie Mellon Pronouncing Dictionary (CMUdict) - Version 0.7b, November 2014. http://www.speech.cs.cmu.edu/cgi-bin/cmudict, as of November 15, 2021.

[45] Kim Lyons. Amazon is Still Crushing Google and Apple in the Smart Speaker Market, February 2020. https://www.theverge.com/2020/2/10/21131988/, as of November 15, 2021.

[46] Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. Privacy Attitudes of Smart Speaker Users. *Privacy Enhancing Technologies*, 2019(4):250—-271, July 2019.

[47] Sri Harish Mallidi, Roland Maas, Kyle Goehner, Ariya Rastrow, Spyros Matsoukas, and Björn Hoffmeister. Device-Directed Utterance Detection. In *Interspeech*, pages 1225–1228, Hyderabad, India, September 2018. ISCA.

[48] Abraham H. Mhaidli, Manikandan Kandadai Venkatesh, Yixin Zou, and Florian Schaub. Listen Only When Spoken To: Interpersonal Communication Cues as Smart Speaker Privacy Controls. *Privacy Enhancing Technologies*, 2020(2):251–270, April 2020.

[49] Hooman Mohajeri Moghaddam, Gunes Acar, Ben Burgess, Arunesh Mathur, Danny Yuxing Huang, Nick Feamster, Edward W. Felten, Prateek Mittal, and Arvind Narayanan. Watching You Watch: The Tracking Ecosystem of Over-the-Top TV Streaming Devices. In *ACM Conference on Computer and Communications Security*, CCS '19, pages 131–147, London, United Kingdom, November 2019. ACM.

[50] Gonzalo Navarro. A Guided Tour to Approximate String Matching. *ACM Computing Surveys*, 33(1):31–88, March 2001.

[51] Orange, S.A. Orange Launches the Voice Assistant Djingo, November 2019. https://djingo.orange.fr/, as of November 15, 2021.

[52] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR Corpus Based on Public Domain Audio Books. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '15, pages 5206–5210, Brisbane, Queensland, Australia, April 2015. IEEE.

[53] Douglas B. Paul and Janet M. Baker. The Design for the Wall Street Journal-based CSR Corpus. In *Workshop on Speech and Natural Language*, HLT '92, pages 357–362, Harriman, New York, USA, February 1992. Association for Computational Linguistics.

[54] David Pierce. Google's New Magic Number for Storing Personal Data: 18 Months, June 2020. https://www.protocol.com/google-delete-data-18-months, as of November 15, 2021.

[55] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. The Kaldi Speech Recognition Toolkit. In *IEEE Workshop on Automatic Speech Recognition & Understanding*, ASRU '11, pages 1–4, Waikoloa, Hawaii, USA, December 2011. IEEE.

[56] Anirudh Raju, Sankaran Panchapagesan, Xing Liu, Arindam Mandal, and Nikko Strom. Data Augmentation for Robust Keyword Spotting under Playback Interference. *CoRR*, abs/1808.00563:1–6, August 2018.

[57] Mike Rodehorst. Why Alexa Won't Wake Up When She Hears Her Name in Amazon's Super Bowl Ad, January 2019. https://www.amazon.science/blog/why-alexa-wont-wake-up-when-she-hears-her-name-in-amazons-super-bowl-ad, as of November 15, 2021.

[58] Lea Schönherr, Thorsten Eisenhofer, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Imperio: Robust Over-the-Air Adversarial Examples for Automatic Speech Recognition Systems. In *Annual Conference on Computer Security Applications*, ACSAC '20, pages 843–855, Virtual Conference, December 2020. ACM.

[59] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding. In *Symposium on Network and Distributed System Security*, NDSS '19, pages 1–15, San Diego, California, USA, February 2019. ISOC.

[60] Timothy Scott. Smart Speakers Statistics: Report 2021, December 2020. https://speakergy.com/smart-speakers-statistics/, as of November 15, 2021.

[61] Siddharth Sigtia, John Bridle, Hywel Richards, Pascal Clark, Erik Marchi, and Vineet Garg. Progressive Voice Trigger Detection: Accuracy vs. Latency. In *International Conference on Acoustics, Speech and Signal Processing*, ICASSP '21, pages 6843–6847, Virtual Conference, June 2021. IEEE.

[62] Siddharth Sigtia, Pascal Clark, Rob Haynes, Hywel Richards, and John Bridle. Multi-Task Learning for Voice Trigger Detection. In *International Conference on Acoustics, Speech and Signal Processing*, ICASSP '20, pages 7449–7453, Barcelona, Spain, May 2020. IEEE.

[63] Siddharth Sigtia, Erik Marchi, Sachin Kajarekar, Devang Naik, and John Bridle. Multi-Task Learning for Speaker Verification and Voice Trigger Detection. In *International Conference on Acoustics, Speech and Signal Processing*, ICASSP '20, pages 6844–6848, Barcelona, Spain, May 2020. IEEE.

[64] Manish Singh. Amazon's Alexa Now Speaks Hindi, September 2019. https://techcrunch.com/2019/09/18/amazon-alexa-hindi-india/, as of November 15, 2021.

[65] David Snyder, Guoguo Chen, and Daniel Povey. MUSAN: A Music, Speech, and Noise Corpus. *CoRR*, abs/1510.08484:1–4, October 2015.

[66] Madiha Tabassum, Tomasz Kosiundefinedski, Alisa Frik, Nathan Malkin, Primal Wijesekera, Serge Egelman, and Heather Richter Lipford. Investigating Users' Preferences and Expectations for Always-Listening Voice Assistants. *Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(4), December 2019.

[67] Rachael Tatman. Gender and Dialect Bias in YouTube's Automatic Captions. In *ACL Workshop on Ethics in Natural Language Processing*, EthNLP '17, pages 53–59, Valencia, Spain, April 2017. ACL.

[68] United States Census Bureau. American Community Survey – Language Spoken at Home, September 2020. https://www.census.gov/acs/www/about/why-we-ask-each-question/language/, as of November 15, 2021.

[69] Tavish Vaidya, Yuankai Zhang, Micah Sherr, and Clay Shields. Cocaine Noodles: Exploiting the Gap between Human and Machine Speech Recognition. In *USENIX Workshop on Offensive Technologies*, WOOT '15, pages 1–14, Washington, District of Columbia, USA, August 2015. USENIX.

[70] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. In *Speech Synthesis Workshop*, SSW '16, pages 125–125, Sunnyvale, California, USA, September 2016. ISCA.

[71] Lente Van Hee, Denny Baert, Tim Verheyden, and Ruben Van Den Heuvel. Google Employees Are Eavesdropping, Even in Your Living Room, July 2019. https://www.vrt.be/vrtnws/en/2019/07/10/google-employees-are-eavesdropping-even-in-flemish-living-rooms/, as of November 15, 2021.

[72] Chris Welch. Amazon's Alexa Can Now Recognize Different Voices and Give Personalized Responses, October 2017. https://www.theverge.com/circuitbreaker/2017/10/11/16460120/, as of November 15, 2021.

[73] Minhua Wu, Sankaran Panchapagesan, Ming Sun, Jiacheng Gu, Ryan Thomas, Shiv Naga Prasad Vitaladevuni, Björn Hoffmeister, and Arindam Mandal. Monophone-Based Background Modeling for Two-Stage On-Device Wake Word Detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '18, pages 5494–5498, Calgary, Alberta, Canada, April 2018. IEEE.

[74] Eric Zeng, Shrirang Mare, and Franziska Roesner. End User Security and Privacy Concerns with Smart Homes. In *Symposium on Usable Privacy and Security*, SOUPS '17, pages 65–80, Santa Clara, California, USA, July 2017. USENIX.

[75] Nan Zhang, Xianghang Mi, Xuan Feng, XiaoFeng Wang, Yuan Tian, and Feng Qian. Dangerous Skills: Understanding and Mitigating Security Risks of Voice-Controlled Third-Party Functions on Virtual Personal Assistant Systems. In *IEEE Symposium on Security and Privacy*, SP '19, pages 1381–1396, San Francisco, California, USA, May 2019. IEEE.