Skill Squatting Attacks on Amazon Alexa

Deepak Kumar, Riccardo Paccagnella, Paul Murley, Eric Hennenfent, Joshua Mason, Adam Bates, and Michael Bailey, ***University of Illinois, Urbana-Champaign***

https://www.usenix.org/conference/usenixsecurity18/presentation/kumar

This paper is included in the Proceedings of the 27th USENIX Security Symposium.

August 15–17, 2018 • Baltimore, MD, USA

ISBN 978-1-931971-46-1

Open access to the Proceedings of the 27th USENIX Security Symposium is sponsored by USENIX.

Skill Squatting Attacks on Amazon Alexa

Deepak Kumar Riccardo Paccagnella Paul Murley Eric Hennenfent
Joshua Mason Adam Bates Michael Bailey

University of Illinois Urbana-Champaign

Abstract

The proliferation of the Internet of Things has increased reliance on voice-controlled devices to perform everyday tasks. Although these devices rely on accurate speech-recognition for correct functionality, many users experience frequent misinterpretations in normal use. In this work, we conduct an empirical analysis of interpretation errors made by Amazon Alexa, the speech-recognition engine that powers the Amazon Echo family of devices. We leverage a dataset of 11,460 speech samples containing English words spoken by American speakers and identify where Alexa misinterprets the audio inputs, how often, and why. We find that certain misinterpretations appear consistently in repeated trials and are *systematic*. Next, we present and validate a new attack, called *skill squatting*. In skill squatting, an attacker leverages systematic errors to route a user to malicious application without their knowledge. In a variant of the attack we call *spear skill squatting*, we further demonstrate that this attack can be targeted at specific demographic groups. We conclude with a discussion of the security implications of speech interpretation errors, countermeasures, and future work.

1 Introduction

The popularity of commercial Internet-of-Things (IoT) devices has sparked an interest in voice interfaces. In 2017, more than 30 M smart speakers were sold [10], all of which use voice as their primary control interface [28]. Voice interfaces can be used to perform a wide array of tasks, such as calling a cab [11], initiating a bank transfer [2], or changing the temperature inside a home [8].

In spite of the growing importance of speech-recognition systems, little attention has been paid to their shortcomings. While the accuracy of these systems is improving [37], many users still experience frequent misinterpretations in everyday use. Those who speak with accents report especially high error rates [36] and other studies report differences in the accuracy of voice-recognition systems when operated by male or female voices [40, 46]. Despite these reports, we are unaware of any independent, public effort to quantify the frequency of speech-recognition errors.

In this work, we conduct an empirical analysis of interpretation errors in speech-recognition systems and investigate their security implications. We focus on Amazon Alexa, the speech-recognition system that powers 70% of the smart speaker market [3], and begin by building a test harness that allows us to utilize Alexa as a black-box transcription service. As test cases, we use the Nationwide Speech Project (NSP) corpus, a dataset of speech samples curated by linguists to study speech patterns [19]. The NSP corpus provides speech samples of 188 words from 60 speakers located in six distinct "dialect-regions" in the United States.

We find that for this dataset of 11,460 utterances, Alexa has an aggregate accuracy rate of 68.9% on single-word queries. Although 56.4% of the observed errors appear to occur unpredictably (i.e., Alexa makes diverse errors for a distinct input word), 12.7% of them are *systematic*—they appear consistently in repeated trials across multiple speakers. As expected, some of these systematic errors (33.3%) are due to words that have the same pronunciation but different spellings (i.e., homophones). However, other systematic errors (41.7%) can be modeled by differences in their underlying phonetic structure.

Given our analysis of misinterpretations in Amazon Alexa, we consider how an adversary could leverage these systematic interpretation errors. To this end, we introduce a new attack, called *skill squatting*, that exploits Alexa misinterpretations to surreptitiously cause users to trigger malicious, third-party skills. Unlike existing work, which focuses on crafting adversarial audio input to inject voice commands [15, 39, 42, 48, 49], our attack exploits intrinsic error within the opaque natural language processing layer of speech-recognition systems and requires an adversary to only register a public skill. We demonstrate

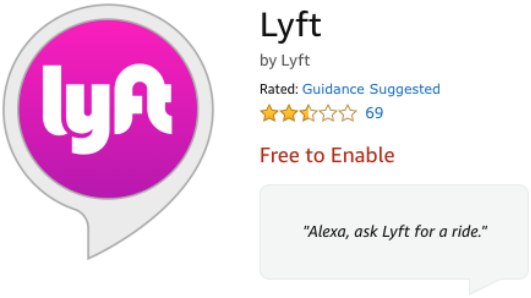


Figure 1: **Example of an Alexa skill** — Alexa skills are applications that can perform useful tasks based on voice input. For example, the Lyft skill [7] allows users to request a ride by saying "Alexa, ask Lyft for a ride."

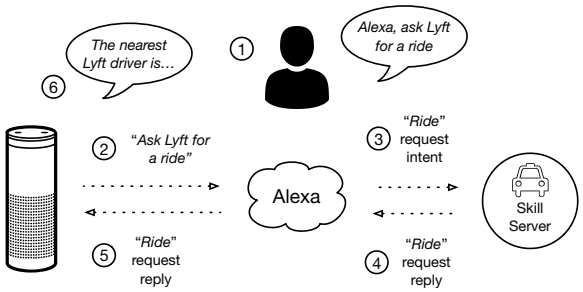


Figure 2: **User-skill interaction in Alexa** — A typical user interaction with an Alexa skill, using an Echo device. In this example, a user interacts with the Lyft skill to request a ride.

this attack in a developer environment and show that we are able to successfully "squat" skills, meaning that Alexa invokes the malicious skill instead of a user-intended target skill at least once for 91.7% of the words that have systematic errors. We then consider how an adversary may improve this attack. To this end, we introduce a variant of skill squatting, called *spear skill squatting*, which exploits systematic errors that uniquely target individuals based on either their dialect-region or their gender. We demonstrate that such an attack is feasible in 72.7% of cases by dialect-region and 83.3% of cases by gender.

Ultimately, we find that an attacker can leverage systematic errors in Amazon Alexa speech-recognition to cause undue harm to users. We conclude with a discussion of countermeasures to our presented attacks. We hope our results will inform the security community about the potential security implications of interpretation errors in voice systems and will provide a foundation for future research in the area.

2 Background

2.1 Voice Interfaces

Voice interfaces are rooted in speech-recognition technology, which has been a topic of research since the 1970s [26]. In recent years, voice interfaces have become a general purpose means of interacting with computers, largely due to the proliferation of the Internet of Things. In many cases, these interfaces entirely supplant traditional controls such as keyboards and touch screens. Smart speakers, like the Amazon Echo and Google Home, use voice interfaces as their primary input source. As of January 2018, an estimated 39 M Americans 18 years or older own a smart speaker [10], the most popular belonging to the Amazon Echo family.

2.2 Amazon Alexa Skills

In this work, we focus on Amazon Alexa [14], the speech-recognition engine that powers the Amazon Echo family of devices, as a state-of-the-art commercial voice interface. In order to add extensibility to the platform, Amazon allows the development of third-party applications, called "skills", that leverage Alexa voice services. Many companies are actively developing Alexa skills to provide easy access to their services through voice. For example, users can now request rides through the Lyft skill (Figure 1) and conduct everyday banking tasks with the American Express skill [4].

Users interact with skills directly through their voice. Figure 2 illustrates a typical interaction. The user first invokes the skill by saying the skill name or its associated invocation phrase (①). The user's request is then routed through Alexa cloud servers (②), which determine where to forward it based on the user input (③). The invoked skill then replies with the desired output (④), which is finally routed from Alexa back to the user (⑤). Up until April of 2017, Alexa required users to enable a skill to their account, in a manner similar to downloading a mobile application onto a personal device. However, Alexa now offers the ability to interact with skills without enabling them [32].

2.3 Phonemes

In this work, we consider how the pronunciation of a word helps explain Alexa misinterpretations. Word pronunciations are uniquely defined by their underlying *phonemes*. Phonemes are a speaker-independent means of describing the units of sound that define the pronunciation of a particular word. In order to enable text-based analysis of English speech, the Advanced Research Projects Agency (ARPA) developed ARPAbet, a set of phonetic transcription codes that represent phonemes of General American English using distinct sequences of ASCII characters [30]. For example, the phonetic representation of

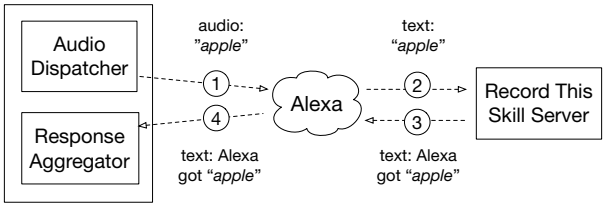


Figure 3: **Speech-to-Text Test Harness Architecture** — By building an experimental skill (called "Record This"), we are able to use the Amazon Alexa speech recognition system as a black box transcription service. In this example, the client sends a speech sample of the word "apple" ①, Alexa transcribes it for the skill server ②, which then returns the transcription as a reply to Alexa ③ and back to the client ④.

the word "pronounce" using the ARPAbet transcription codes is `P R AH N AW N S`. For the scope of this work, we define the phonetic spelling of a word as its ARPAbet phonetic representation, with each ARPAbet character representing a single phoneme. There are 39 phonemes in the ARPAbet. We rely on the CMU Pronunciation Dictionary [22] as our primary source for word to phonemes conversion.

3 Methodology

In this section, we detail the architecture of our test harness, provide an overview of the speech corpora used in our analysis, and explain how we use both to investigate Alexa interpretation errors.

3.1 Speech-to-Text Test Harness

Alexa does not directly provide speech transcriptions of audio files. It does, however, allow third-party skills to receive literal transcriptions of speech as a developer API feature. In order to use Alexa as a transcription service, we built an Alexa skill (called "Record this") that records the raw transcript of input speech. We then developed a client that takes audio files as input and sends them through the Alexa cloud to our skill server. In order to start a session with our Alexa skill server, the client first sends an initialization command that contains the name of our custom skill. Amazon then routes all future requests for that session directly to our "Record this" skill server. Second, the client takes a collection of audio files as input, batches them, and sends them to our skill server, generating one query per file. We limit queries to a maximum of 400 per minute in order to avoid overloading Amazon's production servers. In addition, if a request is denied or no response is returned, we try up to five times before marking the query as a failure.

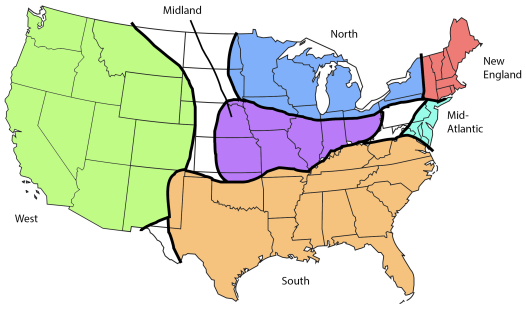


Figure 4: **Dialect-Regions in the U.S.** — Labov et al.'s [31] six dialect regions define broad classes of speech patterns in the United States, which are used to segment Nationwide Speech Project dataset.

Data Source	Speakers	Words	Samples
NSP	60	188	11,460
Forvo	4,990	59,403	91,843

Table 1: **Speech Sources** — We utilize two speech databases, the Nationwide Speech Project (NSP) and Forvo, to aid in our analysis of Alexa misinterpretations. We use the NSP dataset as our primary source for speech samples and the Forvo dataset solely for cross-validation.

Figure 3 illustrates this architecture. For each audio file sent from the client (①), Alexa sends a request to our skill server containing the understood text transcription (②). The server then responds with that same transcription (③) through the Alexa service back to the client (④). The client aggregates the transcriptions in a results file that maps input words to their output words for each audio sample.

3.2 Speech Corpora

In order to study interpretation errors in Alexa, we rely on two externally collected speech corpora. A full breakdown of these datasets is provided in Table 1.

NSP The Nationwide Speech Project (NSP) is an effort led by Ohio State University to provide structured speech data from a range of speakers across the United States [19]. The NSP corpus provides speech from a total of 60 speakers from six geographical "dialect-regions", as defined by Labov et al. [31]. Figure 4 shows each of these speech regions — Mid-Atlantic, Midland, New England, North, South, and West — over a map of the United States. In particular, five male and five female speakers from each region provide a set of 188 single-word recordings, 76 of which are single-syllable words (e.g. "mice", "dome", "bait") and 112 are multi-syllable words (e.g. "alfalfa", "nectarine"). These single-word files provide a

total of 11,460 speech samples for further analysis and serve as our primary source of speech data. In addition, NSP provides metadata on each speaker, including gender, age, race, and hometown.

Forvo We also collect speech samples from the Forvo website [6], which is a crowdsourced collection of pronunciations of English words. We crawled `forvo.com` for all audio files published by speakers in the United States, on November 22nd, 2017. This dataset contains 91,843 speech samples covering 59,403 words from 4,991 speakers. Unfortunately, the Forvo data is non-uniform and sparse. 40,582 (68.3%) of the words in the dataset are only spoken by a single speaker, which makes reasoning about interpretation errors in such words difficult. In addition, the audio quality of each sample varies from speaker to speaker, which adds difficult-to-quantify noise in our measurements. In light of these observations, we limit our use of these data to only cross-validation of our results drawn from NSP data.

3.3 Querying Alexa

We use our test harness to query Alexa for a transcription of each speech sample in the NSP dataset. First, we observe that Alexa does not consistently return the same transcription when processing the same speech sample. In other words, Alexa is non-deterministic, even when presented with identical audio files over reliable network communication (i.e., TCP). This may be due to some combination of A/B testing, system load, or evolving models in the Alexa speech-recognition system. Since we choose to treat Alexa as a black box, investigating this phenomenon is outside the scope of this work. However, we note that this non-determinism will lead to unavoidable variance in our results. To account for this variance, we query each audio sample 50 times. This provides us with 573,000 data points across 60 speakers. Over all these queries, Alexa did not return a response on 681 (0.1%) of the queries, which we exclude from our analysis. We collected this dataset of 572,319 Alexa transcriptions on January 14th, 2018 over a period of 24 hours.

3.4 Scraping Alexa Skills

Part of our analysis includes investigating how interpretation errors relate to Alexa skill names. We used a third-party aggregation database [1] to gather a list of all the skill names that were publicly available on the Alexa skills store. This list contains 25,150 skill names, of which 23,368 are unique. This list was collected on December 27th, 2017.

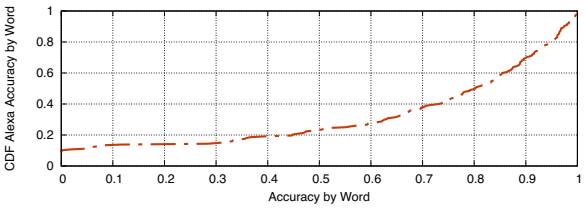


Figure 5: **Word Accuracy** — The accuracy of Alexa interpretations by word is shown as a cumulative distribution function. 9% of the words in our dataset are never interpreted correctly and 2% are always interpreted correctly. This shows substantial variance in misinterpretation rate among words.

3.5 Ethical Considerations

Although we use speech samples collected from human subjects, we never interact with subjects during the course of this work. We use public datasets and ensure our usage is in line with their provider's terms of service. All requests to Alexa are throttled so to not affect the availability of production services. For all attacks presented in this paper, we test them only in a controlled, developer environment. Furthermore, we do not attempt to publish a malicious skill to the public skill store. We have disclosed these attacks to Amazon and will work with them through the standard disclosure process.

4 Understanding Alexa Errors

In this section, we conduct an empirical analysis of the Alexa speech-recognition system. Specifically, we measure its accuracy, quantify the frequency of its interpretation errors, classify these errors, and explain why such errors occur.

4.1 Quantifying Errors

We begin our analysis by investigating how well Alexa transcribes the words in our dataset. We find that Alexa successfully interprets only 394,715 (68.9%) out of the 572,319 queries.

In investigating where Alexa makes interpretation errors, we find that errors do not affect all words equally. Figure 5 shows the interpretation accuracy by individual words in our dataset. Only three words (2%) are always interpreted correctly. In contrast, 9% of words are always interpreted *incorrectly*, indicating that Alexa is poor at correctly interpreting some classes of words. Table 2 characterizes these extremes by showing the top 10 misinterpreted words as well as the top 10 correctly interpreted words in our dataset. We find that words with the lowest accuracy tend to be small, single-syllable words, such as "bean", "calm", and "coal". Words with the highest

Word	Accuracy
Bean	0.0%
Calm	0.0%
Coal	0.0%
Con	0.0%
Cot	0.0%
Dock	0.0%
Heal	0.0%
Lull	0.0%
Lung	0.0%
Main	0.0%

(a) Lowest Accuracy Rate

Word	Accuracy
Forecast	100.0%
Robin	100.0%
Tiger	100.0%
Good	99.9%
Happily	99.8%
Dandelion	99.7%
Serenade	99.6%
Liberator	99.3%
Circumstance	99.3%
Paragraph	99.3%

(b) Highest Accuracy Rate

Table 2: **Words with Highest and Lowest Accuracy** — The best and worst interpretation accuracies for individual words are shown here. We find that the words with the lowest accuracy seem to be small, single syllable words.

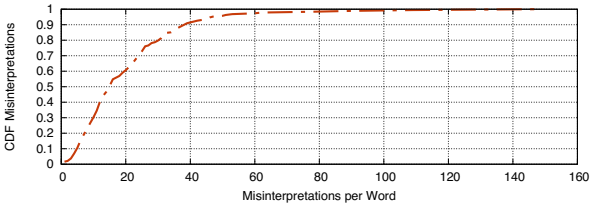


Figure 6: **Unique Misinterpretations per Word** — The number of unique misinterpretations per word is shown as a cumulative distribution function. Even among words that are poorly understood by Alexa, there is variance in the number of unique misinterpretations. The median number of unique misinterpretations is 15, with a heavy tail. In the worst case, the word "unadvised" is misinterpreted in 147 different ways by Alexa.

accuracy are mixed. Many of the top words contain two or three syllables, such as "forecast" and "robin". In one counter example, the word "good" was interpreted correctly 99.9% of the time.

4.2 Classifying Errors

Even among words that are poorly understood by Alexa, there is significant variance in the number of unique misinterpretations. For example, the word "bean" has a 0% accuracy rate and is misinterpreted in 12 different ways, such as "been", "beam", and "bing". In contrast, the word "unadvised" was also never interpreted correctly, but misinterpreted in 147 different ways, such as "an advised", "i devised", and "hundred biased". Figure 6 shows the number of unique misinterpretations per word. The median number of misinterpretations is 15, but with a heavy tail.

In investigating the distributions of misinterpretations per word, we observe that, for each of the 188 words, there are one or two interpretations that Alexa outputs more frequently than the others. Motivated by this observation, we introduce the notion of the "most common error" (MCE) for a given word. As an example, consider the word "boil", which is misinterpreted 100% of the time. The MCE of "boil" is the word "boyle", which accounts for 94.3% (MCE Rate) of the errors. In this sense, the rate at which the MCE occurs serves as a measure of how random the distribution of misinterpretations is. Because "boyle" accounts for the majority of its interpretation errors, we thus claim that "boil" has a predictable misinterpretation distribution.

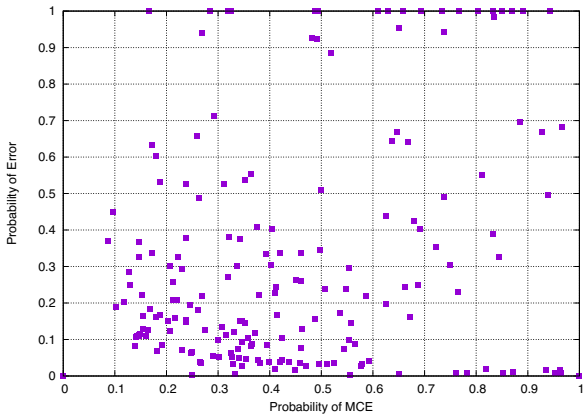


Figure 7: **Error Rate vs MCE** — We plot the error rate by the rate of the most common error for all the words of our dataset. Points in the upper right quadrant represent words that are misinterpreted both frequently and consistently. In our dataset of 188 words, 24 (12.8%) fall in the upper right quadrant.

To visualize the rate and randomness of interpretation errors per word, we plot the error rate for each word along with its MCE rate (Figure 7). This graphical representation provides us with a clearer picture of interpretation errors in Alexa. We then split this plot into four quadrants — quadrant I (upper-right), II (upper-left), III (bottom-left), and IV (bottom-right).

The majority (56.4%) of words in our dataset fall into quadrant III (bottom-left). These are words that are both interpreted correctly most of the time and do not have a prevalent MCE. Instead, they have uncommon errors with no obvious pattern. 21.3% of words appear in quadrant IV (bottom-right). These are words that are often interpreted correctly, but do have a prevalent MCE. There are 9.6% of the words in our dataset that appear in quadrant II (top-left), meaning they are misinterpreted often, but do not feature a prevalent MCE. These are likely to be words that Alexa is poor at understanding altogether. As an example, the word "unadvised", which has 147 unique misinterpretations, appears in this quadrant. The final class of words, in quadrant I (upper-right), are those that are misinterpreted more than 50% of the time and have an MCE that appears in more than 50% of the errors. These are words that are Alexa misunderstands both frequently

Word	MCE	Word Phonemes	MCE Phonemes
rip	rap	R IH P	R AE P
lung	lang	L AH NG	L AE NG
wet	what	W EH T	W AH T
dime	time	D AY M	T AY M
bean	been	B IY N	B IH N
dull	doll	D AH L	D AA L
coal	call	K OW L	K AO L
luck	lock	L AH K	L AA K
loud	louder	L AW D	L AW D ER
sweeten	Sweden	S W IY T AH N	S W IY D AH N

Table 3: **Phonetic Structure of Systematic Errors** — We show the underlying phonetic structure of the ten systematic errors that seem to appear due to Alexa confusing certain phonemes with others. In each case, the resultant MCE is at an edit distance of just one phoneme from the intended word.

and in a consistent manner. There are 24 (12.8%) such words in our dataset.

4.3 Explaining Errors

We now have a classification for interpretation errors from our dataset. Moreover, we identified 24 words for which Alexa consistently outputs one wrong interpretation. We next investigate why these *systematic* errors occur.

Homophones Unsurprisingly, eight (33.3%) of these errors, including "sail" to "sale", "calm" to "com", and "sell" to "cell" are attributable to the fact that these words are homophones, as they have the same pronunciation, but different spellings. Of these, five are cases where Alexa returns a proper noun (of a person, state, band or company) that is a homophone with the spoken word, for example, "main" to "Maine", "boil" to "Boyle", and "outshine" to "Outshyne".

Compound Words Two (8.3%) other systematic errors occur due to compound words. Alexa appears to break these into their constituent words, rather than return the continuous compound word. For example, "superhighway" is split into "super highway" and "outdoors" is split into "out doors".

Phonetic Confusion Ten (41.7%) of the systematic errors can be explained by examining the underlying phonetic structures of the input words and their errors: in each case, the MCE differs from the spoken word by just a single phoneme. For example, the MCE for the word "wet" is the word "what". The phonetic spelling of "wet" is `W EH T`, whereas the phonetic spelling of "what" is `W AH T`. These errors show that Alexa often misunderstands certain specific phonemes within words while correctly interpreting the rest of them. A full list of the phonetic structures for these cases is shown in Table 3.

Other Errors We could not easily explain three (12.5%) of the errors: "mill" to "no", "full" to "four" and "earthy" to "Fi". Even in listening to each speech sample individually, we found no auditory reason why this interpretation error occurs. One surprising error ("preferably" to "preferrably") occurred because Alexa returned a common misspelling of the intended word. This may be caused by a bug in the Alexa system itself.

5 Skill Squatting

Our empirical analysis uncovers the existence of frequently occurring, predictable errors in Amazon Alexa. We next investigate how an adversary can leverage these errors to cause harm to users in the Alexa ecosystem. To this end, we introduce a new attack, called *skill squatting*, which exploits predictable errors to surreptitiously route users to a malicious Alexa skill. The core idea is simple — given a systematic error from one word to another, an adversary constructs a malicious skill that has a high likelihood of confusion with a target skill on the Alexa skills store. When a user attempts to access a desired skill using their voice, they are routed instead to the malicious skill, due to a systematic error in the interpretation of the input. This attack is most similar in style to domain name typosquatting, where an attacker predicts a common "typo" in domain names and abuses it to hijack a request [35, 43, 44, 45]. However, typosquatting relies on the user to make a mistake when typing a domain; in contrast, our attack is intrinsic to the speech-recognition service itself. In this section, we evaluate the skill squatting attack and explore what it looks like in the wild.

5.1 Will This Attack Work End-To-End?

Up to this point, our model of interpretation errors has been entirely constructed based on observations outside of a skill invocation environment. We next investigate whether these errors can be exploited in a skill invocation environment, to redirect the processing of an Alexa query to an attacker-controlled skill server.

Our testing process is as follows: given a model of predictable errors, we build pairs of skills with names that are frequently confused by Alexa. For example, because "boil" is frequently confused with "boyle", we would build two skills: one with the name `Boil` and one with the name `Boyle`. We call these skills the *target skill* (or *squattable skill*) and the *squatted skill*. We refer to words with these predictable, frequently occurring errors as *squattable*. If an attack is successful, Alexa will trigger the squatted skill when a request for the target skill is received. For example, when a user says:

> "Alexa, ask `Boil` hello."

Target Skill	Squatted Skill	Success Rate	Target Skill	Squatted Skill	Success Rate
Coal	Call	100.0%	Dime	Time	65.2%
Lung	Lang	100.0%	Wet	What	62.1%
Sell	Cell	100.0%	Sweeten	Sweden	57.4%
Heal	He'll	96.4%	Earthy	Fi	53.3%
Sail	Sale	95.0%	Full	Four	26.8%
Accelerate	Xcelerate	93.7%	Outshine	Outshyne	21.2%
Rip	Rap	88.8%	Superhighway	Super Highway	19.7%
Mill	No	84.6%	Meal	Meow	18.3%
Con	Khan	84.2%	Bean	Been	17.8%
Luck	Lock	81.9%	Tube	Two	16.7%
Lull	Lol	81.9%	Main	Maine	3.1%
Dull	Doll	80.8%	Boil	Boyle	0.0%
Outdoors	Out Doors	71.0%	Loud	Louder	0.0%
Calm	Com	67.9%			

Table 4: **Skill Squatting Validation** — We show the results of testing 27 skill squatting attacks. The pairs of target and squatted skills are built using the squattable words of our training set. The success rates are computed by querying the speech samples of our test set. We are able to successfully squat 25 (92.6%) of the skills at least one time, demonstrating the feasibility of the attack.

They will instead be routed to the `Boyle` skill.

In order to demonstrate that our attack will work on speakers we have not previously seen, we use two-fold cross validation over the 60 speakers in our dataset. We divide the set randomly into two halves, with 30 speakers in each half. We build an error model using the first half of the speakers (training set) and then use this model to build pairs of target and squatted skills. The analysis of this training set results in 27 squattable words, all of which are detailed in Table 4. For each speaker in the test set, we construct a request to each of the 27 target skills and measure how many times the squatted skill is triggered. We repeat this process five times to address non-determinism in Alexa responses. As an ethical consideration, we test our attack by registering our skills in a developer environment and not on the public Alexa skills store, to avoid the possibility of regular users inadvertently triggering them.

Table 4 shows the results of our validation experiment. We are able to successfully squat skills at least once for 25 (92.6%) of the 27 squattable skills. There are two cases in which our squatting attack never works. In the first case, we expect the skill name `loud` to be incorrectly interpreted as the word `louder`. However, because `louder` is a native Alexa command which causes Alexa to increase the volume on the end-user device, when the target is misinterpreted, it is instead used to perform a native Alexa function. We found no clear explanation for the second pair of skills, `Boil/Boyle`.

In other cases, we find that testing the attack in a skill environment results in a very high rate of success. In the `Coal/Call` and `Sell/Cell` pairs, the attack works 100% of the time. We speculate that this is a result of a smaller solution space when Alexa is choosing between skills as opposed to when it is transcribing arbitrary speech within a skill. Ultimately, Table 4 demonstrates that skill squatting attacks are feasible.

Skill	Squatted Skill
Boil an Egg	Boyle an Egg
Main Site Workout	Maine Site Workout
Quick Calm	Quick Com
Bean Stock	Been Stock
Test Your Luck	Test Your Lock
Comic Con Dates	Comic Khan Dates
Mill Valley Guide	No Valley Guide
Full Moon	Four Moon
Way Loud	Way Louder
Upstate Outdoors	Upstate Out
Rip Ride Rockit	Rap Ride Rocket

Table 5: **Squattable Skills in the Alexa skills store** — We show 11 examples of squattable skills publicly available in the Alexa skill store, as well as squatted skill names an attacker could use to "squat" them.

5.2 Squatting Existing Skills

We next investigate how an adversary can craft maliciously named skills targeting existing skills in the Alexa skills store, by leveraging the squattable words we identified in Section 4. To this goal, we utilize our dataset of Alexa skill names described in Section 3. First, we split each skill name into its individual words. If a word in a skill exists in our spoken dataset of 188 words, we check whether that word is squattable. If it is, we exchange that word with its most common error to create a new skill name. As an example, the word "calm" is systematically misinterpreted as "com" in our dataset. Therefore, a skill

with the word "calm" can be squatted by using the word "com" in its place (e.g. "quick com" squats the existing Alexa skill "quick calm").

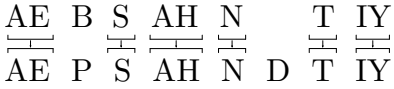
Using the 24 squattable words we identified in Section 4, we find that we can target 31 skill names that currently exist on the Alexa Store. Only 11 (45.8%) of the squattable words appear in Alexa skill names. Table 5 shows one example of a squattable skill for each of these 11 words. We note that the number of squattable skills we identify is primarily limited by the size of our dataset and it is not a ceiling for the pervasiveness of this vulnerability in the Amazon market. To address this shortcoming, in the remainder of this section we demonstrate how an attacker with a limited speech corpus can predict squattable skills using previously-unobserved words.

5.3 Extending The Squatting Attack

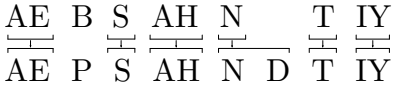
An adversary that attempts this attack using the techniques described thus far would be severely restricted by the size and diversity of their speech corpus. Without many recordings of a target word from a variety of speakers, they would be unable to reliably identify systematic misinterpretations of that word. Considering that many popular skill names make use of novel words (e.g., WeMo) or words that appear less frequently in discourse (e.g., Uber), acquiring such a speech corpus may prove prohibitively costly and, in some cases, infeasible. We now consider how an attacker could amplify the value of their speech corpus by reasoning about Alexa misinterpretations at the phonetic level. To demonstrate this approach, we consider the misinterpretation of "luck" in Table 4."Luck" (`L AH K`) is frequently misinterpreted as "lock" (`L AA K`), suggesting that Alexa experiences confusion specifically between the phonemes `AH` and `AA`. As such, an attacker might predict confusion in other words with the `AH` phoneme (e.g., "duck" to "dock", "cluck" to "clock") without having directly observed those words in their speech corpus.

Unfortunately, mapping an input word's phonemes to a misinterpreted output word's phonemes is non-trivial. The phonetic spelling of the input and output words may be of different lengths, creating ambiguity in the attribution of an error to each input phoneme. Consider the following example from our tests, where the input word "absentee" (`AE, B, S, AH, N, T, IY`) is understood by Alexa as "apps and t." (`AE, P, S, AH, N, D, T, IY`). Moving from left to right, `AE` is correctly interpreted and an input of `B` maps to an output of `P`. However, determining which input phoneme is at fault for the `D` of the output is less clear. In order to attribute errors at the phonetic level, we thus propose a conservative approach that *a)* minimizes the total number of errors attributed and *b)* discards errors that cannot be attributed to a single input phoneme. Our algorithm works in the following steps:

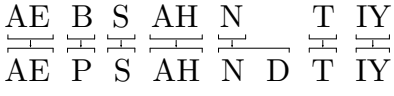
1. We begin by identifying the input-to-output mapping of correct phonemes whose alignment provides the smallest cost (i.e., fewest errors):



2. Based on this alignment, we inspect any additional phonemes inserted into the output that do not correspond to a phoneme in the input. We choose to attribute these output phonemes to a misinterpretation of the phoneme that immediately precedes them in the input. We extend the mappings created in the previous step to include these errors. In our example, we attribute the `D` output phoneme to the `N` input phoneme, mapping `N` to `N D`:



3. Finally, we analyze the remaining unmatched phonemes of the input. We consider unambiguous cases to be where a single phoneme of the input: *a)* occurs between two already mapped pairs of phonemes or is the first or the last phoneme of the input, and *b)* was either omitted (maps to an empty phoneme) or confused with one or two other phonemes in the output. In the example above, we map the phoneme `B` of the input to its single-phoneme misinterpretation as `P` in the output.



We note that this step only attributes an error when its source is unambiguous. There exist some cases where we cannot safely attribute errors and thus we choose to discard an apparent phoneme error. Taking an example from our tests, when the input word "consume" (`K AH N S UW M`) is confused by Alexa as "film" (`F IH L M`), the word error may have happened for reasons unrelated to phoneme misinterpretations and it is not clear how to align input and output except for the final `M` phoneme in both of the words. Since the other phonemes could instead be mapped in many ways, we discard them.

We use this algorithm to create a phoneme error model which provides a mapping from input phonemes to many

possible output phonemes. We next evaluate whether such phoneme error model, built using the NSP dataset, can predict Alexa interpretation errors for words that do not appear in our dataset. To accomplish this, we leverage the Forvo dataset, described in Section 3, as a test set.

First, we exclude from our test set all the speech samples of words that are also in the NSP dataset, since we seek to predict errors for words that we have not used before. Then, we decompose each remaining Forvo word, w, into its phonetic spelling. For every phoneme p in each phonetic spelling we attempt to replace p with each of its possible misinterpretations p_i present in our phoneme error model. We then check if the resultant phoneme string represents an English word, w'. If it does, we mark w' as a potential misinterpretation of w. As an example, consider the word "should", whose phonetic representation is `SH UH D`. The `UH` phoneme is confused with the `OW` phoneme in our phoneme error model, so we attempt a phoneme level swap and get the phoneme string `SH OW D`. This phoneme string maps back to the English word "showed". Thus, we predict that the word "should" will be misinterpreted by Alexa as "showed".

Using this technique, we are able to make error predictions for 12,869 unique Forvo words. To validate the correctness of our predictions, we next collect the actual Alexa interpretations of this set of words. We query each speech sample from this set 50 times using our test harness and record their interpretations. We then check whether any observed interpretation errors in this set are in our predictions. We observe that our predictions are correct for 3,606 (28.8%) of the words in our set. This set is 17.5x larger than our seed of 188 words. This indicates that by extending our word model with a phoneme model, we can successfully predict misinterpretations for a subset of words that we have not previously seen, thus improving the potency of this attack even with a small speech dataset.

5.4 Identifying Existing Confused Skills

We next apply our method of extending our seed-set of errors to identify already existing instances of confused skills in the Alexa skills store. In total, we find 381 unique skill pairs that exhibit phoneme confusion. The largest single contributor is the word "fact", which is commonly misinterpreted as "facts", and "fax". Given the large number of fact-related skills available on the skill store, it is unsurprising that many of these exist in the wild.

In order to determine whether these similarities are due to chance, we investigate each pair individually on the skill store. We find eight examples of squatted skills that we mark as worth investigating more closely (Table 6). We cannot speak to the intention of the skill creators. However, we find it interesting that such examples currently exist on the store. For example, "cat facts" has a corresponding squatted skill, "cat fax", which seemingly performs the same function, though published by a different developer. In another example, "Phish Geek" [9], which purports to give facts about the American rock band Phish, is squatted by "Fish Geek" [5], which gives facts about fish. Anecdotally, one user of "Phish Geek" appears to have experienced squatting, writing in a review:

> I would love it if this actually gave facts about the band. But instead, it tells you things like "Some fish have fangs!"

Ultimately, we have no clear evidence that any of these skills of interest were squatted intentionally. However, this does provide interesting insight into some examples of what an attacker may do and further validates our assertion that our phoneme-based approach can prove useful in finding such examples in the wild.

Skill A	Skill B
Cat Fats	Cat Facts
Pie Number Facts	Pi Number Facts
Cat Facts	Cat Fax
Magic Hate Ball	Magic Eight Ball
Flite Facts	Flight Facts
Smart Homy	Smart Home
Phish Geek	Fish Geek
Snek Helper	Snake Helper

Table 6: **Squatted Skills in the Alexa skills store** — We show examples of squatted skills in the Alexa skills store that drew our attention during manual analysis. Notably, a customer review of the "phish geek" skill noted they were unable to use the application due to common confusion with the "fish geek" skill.

6 Spear Skill Squatting

We have thus far demonstrated skill squatting attacks that target speakers at an aggregate level. We next ask the question, "Can an attacker use skill squatting to target specific groups of people?" To accomplish this, we introduce a variant of the skill squatting attack, called *spear skill squatting*. Spear skill squatting extends skill squatting attacks by leveraging words that only squattable in targeted users' demographic. Spear skill squatting draws its name from the closely related spear phishing family of attacks, which are phishing attacks targeted at specific groups of individuals [25]. In this section, we identify and validate spear skill squatting attacks by targeting speakers based on their geographic region and their gender.

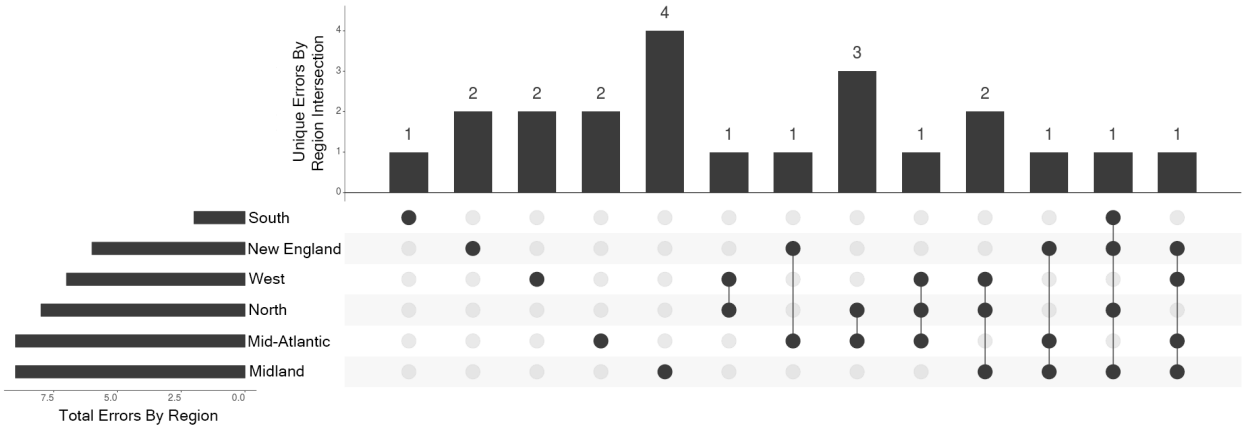


Figure 8: **Regional Intersection of Squattable Words** — We show the 6-way intersection of squattable words by region. Squattable words that affect all regions are omitted. Each region is denoted by a dot in the bottom half of the graph. If a squattable word is shared between two or more regions, the region-dots are connected with a line. The height of each bar corresponds to the number of squattable words per region-intersection. There are 11 squattable words that target just one specific region.

6.1 Demographic Effect on Accuracy

The 60 speakers in the NSP corpus are separated both by dialect-region (10 speakers per region) and gender (30 speakers identify as male, 30 identify as female). We first examine if user demographics play a factor in Alexa accuracy rates.

In order to quantify the differences in accuracy between regions, we run a chi-squared "goodness-of-fit" test. This test is used to determine whether a particular distribution follows an expected distribution. To not over report this statistic given our sample size, we only consider the most common interpretation per speaker per word, rather than use 50 interpretations per speaker per word. As we would like to measure whether interpretation errors happen across all regions with equal probability, our null hypothesis is that there is no significant difference in accuracy between the regions. Our chi-squared test returns a p-value of $6.54 * 10^{-139}$, indicating strong evidence to reject the null hypothesis. This demonstrates that at least one region has a significant difference in accuracy from the rest, with a confidence interval $> 99\%$.

We next investigate whether Alexa has different accuracy rates when interpreting speakers of different genders. We find that Alexa is more accurate when interpreting women (71.9%) than men (66.6%). In addition, a two proportion z-test between the groups shows a statistically significant difference at a confidence interval of 99% (p-value: $1.03 * 10^{-9}$).

6.2 Squattable Words by Demographic

These results indicate that Alexa interprets speakers differently based on their region and their gender. We next investigate whether the interpretation errors for each demographic are systematic and, as a result, can be used by an adversary to launch a spear skill squatting attack.

To identify squattable words based on region, we first split our speakers into their respective dialect-region. Using the techniques outlined in Section 4, we identify the systematic errors that affect each region in isolation. This produces a total of 46 unique squattable words that are occur at least in one region. However, this also includes squattable words that affect every region. Because this attack focuses on targeting specific groups of individuals, we exclude squattable words that affect all regions. After removing these, we are left with 22 squattable words that target a strict subset of all regions. For example, the interpretation error from `Pull/Pole`, only affects systematically speakers from the West, New England, Midland, and Mid-Atlantic regions, but not speakers from the North or South. In contrast, the error `Pal/Pow` only systematically impacts speakers from the Midland region.

Figure 8 shows the distribution of these squattable words per region-intersection. Notably, there are 11 squattable words that each affect *one* region in isolation. Table 7a further breaks down these specific squattable words and their systematic interpretation errors by region. An attacker can leverage any of these in order to target speakers from one specific region.

We then apply the same technique to find squattable words based on speaker gender and observe a similar result — there are squattable words that only affect speakers based on their gender. Table 7b provides a breakdown of the pairs of squattable words and their interpretation errors that affect speakers by gender. There are 12 squattable words that an adversary could leverage to target speakers based on their gender.

Squatted Word	Region	Target Success	Overall Success	Significant?
Tool/Two	South	34.0%	14.1%	✓ (< 0.01)
Dock/Doc	West	97.4%	81.6%	✗ (0.36)
Mighty/My T.	West	20.0%	4.1%	✓ (< 0.01)
Exterior/Xterior	New England	42.9%	22.5%	✓ (0.028)
Meal/Meow	New England	55.6%	34.3%	✓ (< 0.01)
Wool/Well	Midland	50.0%	32.4%	✗ (0.055)
Pal/Pow	Midland	65.9%	37.7%	✓ (< 0.01)
Accuser/Who's There	Midland	26.0%	4.9%	✓ (< 0.01)
Pin/Pen	Midland	26.3%	10.0%	✓ (< 0.01)
Malfunction/No Function	Mid-Atlantic	36.0%	27.5%	✗ (0.23)
Fade/Feed	Mid-Atlantic	59.0%	14.7%	✓ (< 0.01)

(a) Spear Skill Squatting by region

Squatted Word	Gender	Target Success	Overall Success	Significant?
Full/Four	Male	51.1%	11.8%	✓ (< 0.01)
Towel/Tell	Male	83.8%	46.6%	✓ (< 0.01)
Heal/He'll	Male	44.4%	34.9%	✗ (0.26)
Lull/Lol	Male	67.6%	72.4%	✗ (0.45)
Exterior/Xterior	Male	50.0%	30.3%	✓ (< 0.01)
Tube/Two	Male	34.7%	16.8%	✓ (< 0.01)
Preferably/Preferrably	Female	67.6%	36.3%	✓ (< 0.01)
Pull/Paul	Female	75.7%	59.4%	✓ (< 0.01)
Outdoors/Out Doors	Female	69.5%	41.5%	✓ (< 0.01)
Rip/Rap	Female	97.9%	66.7%	✓ (< 0.01)
Hill/Hello	Female	66.0%	28.1%	✓ (< 0.01)
Bull/Ball	Female	39.3%	19.5%	✓ (< 0.01)

(b) Spear Skill Squatting by gender

Table 7: **Validating the Spear Skill Squatting Attack** — We test our spear skill squatting attacks in a developer environment. The last column shows the p-value of a proportion z-test checking whether there is a statistically significant difference, at a confidence interval of 95%, between the success rates of the attack against the region/gender group and the overall population. Our attacks are successful in impacting specific demographic groups 8 out of 11 times by region and 10 out of 12 times by gender.

6.3 Validating Spear Skill Squatting

We next turn to validating that our spear skill squatting attacks will work in a skill environment. To test this, we use a methodology similar to that described in Section 5.1, where we build skills in a developer environment and observe the rate at which our squatted skill is favored over the target skill. Table 7 shows the breakdown of our squatting attempts to target speakers based on both their region and gender. For 8 out of the 11 region-based attacks, we observe a statistically different rate of success for our attack than when compared to the rate of success observed for the rest of the population. Our attack works slightly better when targeting speakers by gender, with an attack working in 10 out of the 12 cases.

Our results provide evidence that such an attack can be successful in a skill environment. We acknowledge that our results are inherently limited in scope by the size of our dataset. An adversary with better knowledge of squattable words can construct new attacks that are outside the purview of our analysis; thus, further scrutiny must be placed on these systems to ensure they do not inadvertently increase risk to the people that use them.

7 Discussion

7.1 Limitations

A core limitation of our analysis is the scope and scale of the dataset we use in our analysis. The NSP dataset only provides 188 words from 60 speakers, which is inadequate for measuring the full scale of systematic misinterpretations of Amazon Alexa. Although our phoneme model extends our observed misinterpretation results to new words, it is also confined by just the errors that appeared from querying the NSP dataset.

Another limitation of our work is that we rely on the key assumption that triggering skills in a development environment works similarly to triggering publicly available skills. However, do not attempt to publish skills

or attack existing skills on the Alexa skills store due to ethical concerns. A comprehensive validation of our attack would require that we work with Amazon to test the skill squatting technique safely in their public, production environment.

7.2 Countermeasures

The skill squatting attack relies on an attacker registering squatted skills. All skills must go through a certification process before they are published. To prevent skill squatting, Amazon could add to the certification process both a word-based and a phoneme-based analysis of a new skill's invocation name in order to determine whether it may be confused with skills that are already registered. As a similar example, domain name registrars commonly restrict the registration of homographs —domains which look very similar visually— of well known domains [34]. These checks seem not to be currently in place on Alexa, as we found 381 pairs of skills with different names, but likely to be squatted on the store (Section 5.4).

Short of pronunciation based attacks, there already exist public skills with identical invocation names on the Alexa skills store. For example, there are currently more than 30 unique skills called "Cat Facts", and the way in which Amazon routes requests in these cases is unclear. Although this is a benign example, it demonstrates that some best practices from other third-party app store environments have not made their way to Alexa yet.

Attacks against targeted user populations based on their demographic information are harder to defend against, as they require a deeper understanding of why such errors occur and how they may appear in the future. Amazon certainly has proprietary models of human speech, likely from many demographic groups. Further analysis is required in order to identify cases in which systematic errors can be used to target a specific population.

8 Future Work

While we have demonstrated the existence of systematic errors and the feasibility of skill squatting attacks, there remain several open challenges to quantifying the scope and scale of these results.

Collecting Richer Datasets. The conclusions we can draw about systematic errors are limited by the size of our speech corpus. We find that, in theory, 16,836 of the 23,238 (72.5%) unique skills in the Alexa skills store could potentially be squatted using our phoneme model. However, without additional speech samples, there is no way for us to validate these potential attacks. In order to more thoroughly investigate systematic errors and their security implications, we must curate a larger, more diverse dataset for future analysis. We suspect that with a larger set of words and speakers, we would not only be able to quantify other systematic errors in Alexa, but also draw stronger conclusions about the role of demographics in speech recognition systems.

Measuring the Harms of Skill Squatting. It remains unclear how effective our attack would be in the wild. In order to observe this, we would need to submit public skills to Amazon for certification. In addition, our work does not explore what an attacker may be able to accomplish once a target skill is successfully squatted. In initial testing, we successfully built phishing attacks on top of skill squatting (for example, against the American Express skill)[1]. However, investigating the scale of such attacks is beyond the scope of this work. We hypothesize that the most significant risk comes from the possibility that an attacker could steal credentials to third party services, but this topic merits further investigation.

Investigating IoT Trust Relationships. On the web, many users have been conditioned to be security conscious, primarily through browser-warnings [13]. However, an outstanding question is whether that conditioning transfers to a voice-controlled IoT setting. If an attacker realizes that users trust voice interfaces more than other forms of computation, they may build better, more targeted attacks on voice-interfaces.

Generalizing our Models. An outstanding question is whether our models can be broadly generalized to other speech-recognition systems. It is unlikely that our Alexa-specific model of systematic errors will translate directly to other systems. However, the techniques we use to build these models will work as long as we can leverage a speech-recognition system as a black box. Future work must be done in replicating our techniques to other speech-recognition systems.

9 Related Work

Our work builds on research from a number of disciplines, including linguistics, the human aspects of security and targeted audio attacks on voice-controlled systems.

Dialects in Speech. Linguists have developed models of English speech since the 1970s, from intonation to rhythm patterns [23]. Recently, researchers have used phoneme and vowel data similar to that of the NSP dataset [19] to study the patterns of speech by region and gender [20, 21, 31]. Clopper has also investigated the effects of dialect variation within sentences on "semantic predictability" —this is the ability of a listener to discern words based on the context in which they appear [18].

[1] `https://youtu.be/kTPkwDzybcc`

Typosquatting and Human Factors. Our work broadly aligns with research about the human aspects of security, such as susceptibility to spam or phishing attacks [25, 27]. Specifically, we focus on a long history of research into domain typosquatting [12, 33, 43, 44]. Using ideas similar to our work, Nikiforakis et al. relied on homophone confusion to find vulnerable domain names [35]. Most recently, Tahir et al. investigated why some URLs are more susceptible to typosquatting than other URLs [45]. Our work also draws on analysis of attack vectors that are beyond simply making mistakes — Kintis et al. studied the longitudinal effects of "combosquatting" attacks, which are variants of typosquatting [29].

Other Skill Squatting Attacks. We are not alone in highlighting the need to investigate the security of speech recognition systems. In a recent preprint, Zhang et al. report a variant of the skill squatting attack based on the observation that Alexa favors the longest matching skill name when processing voice commands [50]. If a user embellished their voice command with naturalistic speech, e.g.,"Alexa, open *Sleep Sounds* please" instead of "Alexa, open *Sleep Sounds*," an attacker may be able to register a skill named *Sleep Sounds please* in order to squat on the user's intended skill. Their attack demonstrates dangerous logic errors in the voice assistant's skills market. In contrast, our work considers more broadly how the intrinsic error present in natural language processing algorithms can be weaponized to attack speech recognition systems.

Audio Attacks. Researchers have shown time after time that acoustic attacks are a viable vector causing harm in computing devices. For example, shooting deliberate audio at a drone can cause it to malfunction and crash [41]. Audio attacks have been used to bias sensor input on Fitbit devices and, further, can manipulate sensor input to fully operate toy RC cars [47]. Audio has also been used as an effective side channel in stealing private key information during key generation [24] and leaking private data through the modification of vibration sensors [38].

Beyond such attacks, several researchers have developed a number of of *adversarial examples* of audio input to trick voice-based interfaces. Carlini et al. demonstrated that audio can be synthesized in a way that is indiscernible to humans, but are actuated on by devices [15]. Further, a number of researchers independently developed adversarial audio attacks that are beyond the range of human hearing [39, 42, 49]. Houdini demonstrated that it is possible to construct adversarial audio files that are not distinguishable from the legitimate ones by a human, but lead to predicted invalid transcriptions by target automatic speech recognition systems [17]. Carlini et al. developed a technique for constructing adversarial audio against Mozilla DeepSpeech with a 100% success rate [16]. More recently, Yuan et al. showed that voice commands can be automatically embedded into songs, while not being detected by a human listener [48].

10 Conclusion

In this work, we investigated the interpretation errors made by Amazon Alexa for 11,460 speech samples taken from 60 speakers. We found that some classes of interpretation errors are *systematic*, meaning they appear consistently in repeated trials. We then showed how an attacker can leverage systematic errors to surreptitiously trigger malicious applications for users in the Alexa ecosystem. Further, we demonstrated how this attack could be extended to target users based on their demographic information. We hope our results inform the security community about the implications of interpretation errors in speech-recognition systems and provide the groundwork for future work in the area.

Acknowledgements

This work was supported in part by the National Science Foundation under contracts CNS 1750024, CNS 1657534, and CNS 1518741. This work was additionally supported by the U.S. Department of Homeland Security contract HSHQDC-17-J-00170. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of their employers or the sponsors.

References

[1] Alexa skills store. https://www.alexaskillstore.com/.

[2] Ally bank. https://www.ally.com/bank/online-banking/how-to-bank-with-ally/alexa/.

[3] Amazon alexa smart speaker market share dips below 70% in u.s., google rises to 25%. https://www.voicebot.ai/2018/01/10/amazon-alexa-smart-speaker-market-share-dips-70-u-s-google-rises-25/.

[4] American express skill. https://www.americanexpress.com/us/content/alexa/.

[5] Fish geek. https://www.amazon.com/Matt-Mitchell-Fish-Geek/dp/B01LMN5RGU/.

[6] Forvo. https://forvo.com/.

[7] Lyft. https://www.amazon.com/Lyft/dp/B01FV34BGE.

[8] Nest thermostat. https://www.amazon.com/Nest-Labs-Inc-Thermostat/dp/B01EIQW9LY.

[9] Phish geek. https://www.amazon.com/EP-Phish-Geek/dp/B01DQG4F0A.

[10] The smart audio report from npr and edison research. http://nationalpublicmedia.com/wp-content/uploads/2018/01/The-Smart-Audio-Report-from-NPR-and-Edison-Research-Fall-Winter-2017.pdf.

[11] Uber. https://www.amazon.com/Uber-Technologies-Inc/dp/B01AYJQ9QK.

[12] P. Agten, W. Joosen, F. Piessens, and N. Nikiforakis. Seven months' worth of mistakes: A longitudinal study of typosquatting abuse. In 22nd Network and Distributed System Security Symposium (NDSS).

[13] D. Akhawe and A. P. Felt. Alice in warningland: A large-scale field study of browser security warning effectiveness. In 22nd USENIX Security Symposium (USENIX).

[14] Amazon. Alexa. https://developer.amazon.com/alexa.

[15] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou. Hidden voice commands. In 25th USENIX Security Symposium (USENIX).

[16] N. Carlini and D. Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In 1st Deep Learning and Security Workshop (DLS).

[17] M. M. Cisse, Y. Adi, N. Neverova, and J. Keshet. Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. In 31st Advances in Neural Information Processing Systems (NIPS).

[18] C. G. Clopper. Effects of dialect variation on the semantic predictability benefit. In Language and Cognitive Processes.

[19] C. G. Clopper. Linguistic experience and the perceptual classification of dialect variation. 2004.

[20] C. G. Clopper and R. Smiljanic. Effects of gender and regional dialect on prosodic patterns in american english. In Journal of Phonetics.

[21] C. G. Clopper and R. Smiljanic. Regional variation in temporal organization in american english. In Journal of Phonetics.

[22] CMU. Cmu pronunciation dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[23] D. Crystal. Prosodic systems and intonation in English. CUP Archive.

[24] D. Genkin, A. Shamir, and E. Tromer. Rsa key extraction via low-bandwidth acoustic cryptanalysis. In 34th International Cryptology Conference (CRYPTO).

[25] G. Ho, A. Sharma, M. Javed, V. Paxson, and D. Wagner. Detecting credential spearphishing in enterprise settings. In 26th USENIX Security Symposium (USENIX).

[26] F. Itakura. Minimum prediction residual principle applied to speech recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing.

[27] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and S. Savage. Spamalytics: An empirical analysis of spam marketing conversion. In 15th ACM conference on Computer and communications security (CCS).

[28] B. Kinsella. 56 million smart speaker sales in 2018 says canalys. https://www.voicebot.ai/2018/01/07/56-million-smart-speaker-sales-2018-says-canalys/.

[29] P. Kintis, N. Miramirkhani, C. Lever, Y. Chen, R. Romero-Gómez, N. Pitropakis, N. Nikiforakis, and M. Antonakakis. Hiding in plain sight: A longitudinal study of combosquatting abuse. In 24th ACM Conference on Computer and Communications Security (CCS), 2017.

[30] A. Klautau. ARPABET and the TIMIT alphabet. https://web.archive.org/web/20160603180727/http://www.laps.ufpa.br/aldebaro/papers/ak_arpabet01.pdf, 2001.

[31] W. Labov, S. Ash, and C. Boberg. The atlas of North American English: Phonetics, phonology and sound change. 2005.

[32] T. Martin. You can now use any alexa skill without enabling it first. https://www.cnet.com/how-to/amazon-echo-you-can-now-use-any-alexa-skill-without-enabling-it-first/.

[33] T. Moore and B. Edelman. Measuring the perpetrators and funders of typosquatting. In 14th International Conference on Financial Cryptography and Data Security.

[34] Namecheap. Do you support idn domains and emoticons? https://www.namecheap.com/support/knowledgebase/article.aspx/238/35/do-you-support-idn-domains-and-emoticons.

[35] N. Nikiforakis, M. Balduzzi, L. Desmet, F. Piessens, and W. Joosen. Soundsquatting: Uncovering the use of homophones in domain squatting. In International Conference on Information Security, 2014.

[36] S. Paul. Voice is the next big platform, unless you have an accent. https://www.wired.com/2017/03/voice-is-the-next-big-platform-unless-you-have-an-accent/.

[37] E. Protalinski. Google's speech recognition technology now has a 4.9% word error rate. https://venturebeat.com/2017/05/17/googles-speech-recognition-technology-now-has-a-4-9-word-error-rate/.

[38] N. Roy. Vibraphone project webpage. http://synrg.csl.illinois.edu/vibraphone/. Last accessed 9 December 2015.

[39] N. Roy, H. Hassanieh, and R. R. Choudhury. Backdoor: Making microphones hear inaudible sounds. In 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys).

[40] ScienceDaily. American Roentgen Ray Society. "Voice Recognition Systems Seem To Make More Errors With Women's Dictation.". https://www.sciencedaily.com/releases/2007/05/070504133050.htm, 2007.

[41] Y. Son, H. Shin, D. Kim, Y. Park, J. Noh, K. Choi, J. Choi, and Y. Kim. Rocking drones with intentional sound noise on gyroscopic sensors. In 24th USENIX Security Symposium (USENIX).

[42] L. Song and P. Mittal. Inaudible voice commands. Preprint, arXiv:1708.07238 [cs.CR], 2017.

[43] J. Spaulding, S. Upadhyaya, and A. Mohaisen. The landscape of domain name typosquatting: Techniques and countermeasures. In 2016 11th International Conference on Availability, Reliability and Security (ARES).

[44] J. Szurdi, B. Kocso, G. Cseh, J. Spring, M. Felegyhazi, and C. Kanich. The long "taile" of typosquatting domain names. In 23rd USENIX Security Symposium (USENIX).

[45] R. Tahir, A. Raza, F. Ahmad, J. Kazi, F. Zaffar, C. Kanich, and M. Caesar. It's all in the name: Why some urls are more vulnerable to typosquatting. In 13th IEEE International Conference on Computer Communications (INFOCOM).

[46] R. Tatman. Google's speech recognition has a gender bias. https://makingnoiseandhearingthings.com/2016/07/12/googles-speech-recognition-has-a-gender-bias/.

[47] T. Trippel, O. Weisse, W. Xu, P. Honeyman, and K. Fu. Walnut: Waging doubt on the integrity of mems accelerometers with acoustic injection attacks. In 2nd IEEE European Symposium on Security and Privacy (Euro S&P).

[48] X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter. Commandersong: A systematic approach for practical adversarial voice recognition. In 27th USENIX Security Symposium (USENIX).

[49] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu. Dolphinattack: Inaudible voice commands. In Proceedings of the ACM Conference on Computer and Communications Security (CCS). ACM, 2017.

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