# Understanding and Mitigating Skill Squatting Attacks in Amazon Alexa

Siva Likitha Valluru
*School of Computing*
*Clemson University*
Clemson, USA
sivaliv@g.clemson.edu

Hruteesh Raja Valluru
*School of Computing*
*Clemson University*
Clemson, USA
hvallur@g.clemson.edu

***Abstract*—Virtual Personal Assistants (VPAs) such as Google Assistant and Amazon Alexa rely on voice input to communicate with users. However, it is widely known that these communication devices lack proper authentication mechanisms, leaving them vulnerable to both known and unknown threats. This project replicates prior research that identified two threats (Voice Squatting Attack (VSA) and Voice Masquerading Attack (VMA)) and develops defenses against these threats. The work of the original authors was recognized by both Google and Amazon in February 2018 and it was shown that the attacks were serious and concerning.**

***Index Terms*—Amazon Alexa, Virtual Personal Assistant (VPA), Voice Squatting Attack (VSA), Voice Masquerading Attack (VMA), Web Crawler, Bi-directional Long Short-Term Memory (BLSTM) Network, Machine Learning, Natural Language Processing (NLP)**

## I. Introduction

Virtual Personal Assistants (VPAs) have become one of the most useful technologies that they have become very commonplace. Especially the VPAs that are integrated to IoT devices which allow the user to interact with them through their voice by saying the commands they want executed instead of touch interaction. This is established with the help of a microphone and a speaker on the device which are used for communication between the user and the device. The assistant does anything that it is programmed to do and is capable of. So it can be used for a wide range of things such as setting the occasional alarm or checking the email. Once it receives a command from a user, it's task is to identify the right command to execute based on the user's speech. This is the main issue with VPAs. They need to understand the command given by the user correctly, otherwise they will be useless or even harmful to the user. If the command given by the user is interpreted differently by the assistant, then it will execute a different command or doesn't do anything at all.

Recently, there are some VPAs that make it possible for the user to get custom functions to execute on the user's command referred to as skills. This enables the assistant to obtain more functions according to the user's needs instead of the user programming it him/herself or unable to do that function through an assistant. These custom, third-party skills cannot be verified by the service provider but it is at the user's judgement to trust that particular skill or not. The problem is that some of these skills may have malicious code that may try to steal information from the user or perform any passive attacks such as eavesdropping. As the virtual assistant is very powerful, it is vulnerable in many ways because it has a microphone that is always listening for a command, so it processes everything heard by the microphone. This is dangerous as the VPA has access to a lot of information about the user since it is personalized.

In this project we will be looking at two common types of attacks that are performed on VPAs that were not known before since they are very subtle. They are Voice Squatting Attack (VSA) and Voice Masquerading Attack (VMA).

The VSA is an attack that exploits the presence of homonyms, words that sound the same but are spelled differently, in voice invoked commands. It also exploits the different phrases that users say to invoke the same command.Since voice based personal assistants function solely due to voice commands, this creates a lot of confusion while processing the command and can be interpreted both ways by the same assistant. There may be a useful skill for which there is a complimentary malicious skill that is present. For example, there may be a skill that sets the alarm when the user gives the command "Set the alarm." But when the user says, "Set the alarm please," then there may be a different malicious skill that is invoked which can be used for running other lines of code.

Voice masquerading attack is basically when there is a malicious skill that poses as a genuine skill but does not perform the function properly. It may report that there is an error and seemingly exited out of the skill but it tries to prolong the interaction in order to keep the microphone listening to the user and sending the information to an attacker. For example if a malicious skill was invoked for the command, "Set the alarm," the VPA may say that it failed and seem like it exited but when the user gives another command, it fakes the interaction by giving the response coded in the malicious skill because it has not yet exited out and the user may think it is the assistant's own response. These are very dangerous attacks as the user may not even be aware of anything malicious going on.

The rest of this proposal is organized as follows: a brief explanation of related work, hypotheses identified, description of the experimental setup, identifying required technology, and giving an anticipated schedule.

## II. Related Work

Prior research has shown that certain attacks can be targeted against specific demographic groups [1]. Based on different dialects spoken across the United States, broad classes of speech patterns were analyzed. By conducting an empirical analysis of Alexa's speech-recognition system, it was shown that variants in pronunciation (measured through phonemes) of certain words/phrases could invoke different skills, thereby showing that the threats, VSA and VMA, are real and serious.

Other works confirmed the potential for hiding malicious voice invocations to VPAs in speech sounds which are perceived by humans as nonsensical. The words 'nonsensical' or 'nonsense' mean the concatenation of words that give no real meaning in context or in everyday usage.

Other authors conducted a Man-in-the-Middle (MITM) attack, exploiting the loopholes in the VPA Skill interfaces of Alexa to redirect the input (user's voice) to a malicious skill, thereby hijacking the conversation. This invasive attack and highly threatening attack is especially predominant in home security systems and smart homes.

An interesting notion that many authors have made is that the scope of their results are ultimately limited by the size of the data sets used. It is always difficult to guarantee that data sets are comprehensive and complete.

## III. Hypotheses

The authors of the original paper have mentioned that they have used web crawlers (discussed in detail in Section IV) to identify all of the skills in Amazon Alexa's Skill Store. Though their research is current, the skills they have managed to acquire may not have been. Newer skills would have been added by not just Amazon but Google as well. This may cause differences in certain statistics.

Additionally, they have used many data sets that originate from Stanford. Even though machine learning and NLP-based algorithms were used to identify suspicious skill recommendations, there does not seem to be a guarantee that Stanford's data sets mimic real-world threats/attacks. Therefore, similar to what the authors stated, it is difficult to know just how much credibility is still intact from 2018.

Finally, last but not least, the authors have not made their findings open-source. Their code is not made public, therefore, their derived performance percentages are estimated to be different from ours.

## IV. Experimental Setup

After our system is implemented, we will test each component against the respective type of attack and use similar techniques to [2] to determine efficacy.

For VSA defense, we plan to identify interesting case studies and potentially make new discoveries by adjusting the thresholds differently than the one proposed in [2].

For VMA defense, the detector requires extensive evaluation to make sure that it is neither overfitting nor underfitting as both pose an equal risk to many VPA users around the world. In [2], the approach they have chosen did not affect their performance at all (latency was negligible). Latency is one of the biggest performance metrics in our project as VPAs are real-time devices and require constant, back-to-back, conversational communication to be considered "effective." Two other evaluation techniques include: measuring effectiveness against prototype attacks and on real-world conversations.

## V. Required Technology

Defense against VSA will be implemented through Java 8 and/or Python 3.x with the use of various web and NLP-related libraries such as BeautifulSoup, URLLib, NumPy, Pandas, NLTK, etc. For defense against VMA, we plan to utilize the free environment provided by Amazon, an Alexa Skill Testing Tool known as *Echosim.io*.

According to the paper, there are roughly 32 categories of Alexa's skills, ranging from Finance to Home to Business, etc. Due to the large number of web pages to extract information from, if needed, we plan to run our implementation in a high-performance computing environment resource such as Clemson University's Palmetto Cluster.

## VI. Anticipated Schedule

We identify four milestones in this project:

- Defense against VSA: Build a web crawler to collect the metadata of all the skills available in Alexa's Skills Market. Later, through the use of *utterance paraphrasing* (finding variations of the same invocation) and *pronunciation comparison* (identifying names with similar pronunciation), we will identify any suspicious skills and report other discoveries.
- Defense against VMA: Build a detector that takes a skill's response and/or the user's utterance as its input to determine whether an impersonation risk is present, and alerts the user once detected. There are two parts to this milestone: building a *Skill Response Checker (SRC)* and *User Intention Classifer (UIC)*.
- Evaluation: One basis of evaluation will include comparing our prototypes with the results of performance metrics and other evaluation techniques (for especially defense against VMA) presented in [2].
- Report: Detail our findings and identify limitations, if any, and other potential vulnerabilities.

Because of our limited practical knowledge with VPAs, IoT devices, and other network-based devices, the work will be split fairly between us for better understanding of the project. Each task's completion and correctness will be ensured by both of us equally. As seen in the schedule down below, tentatively, the physical implementation should be complete in roughly four to five weeks. For the remaining 15 days, we will focus on evaluation, performance, and optimization of the algorithms used in the system.

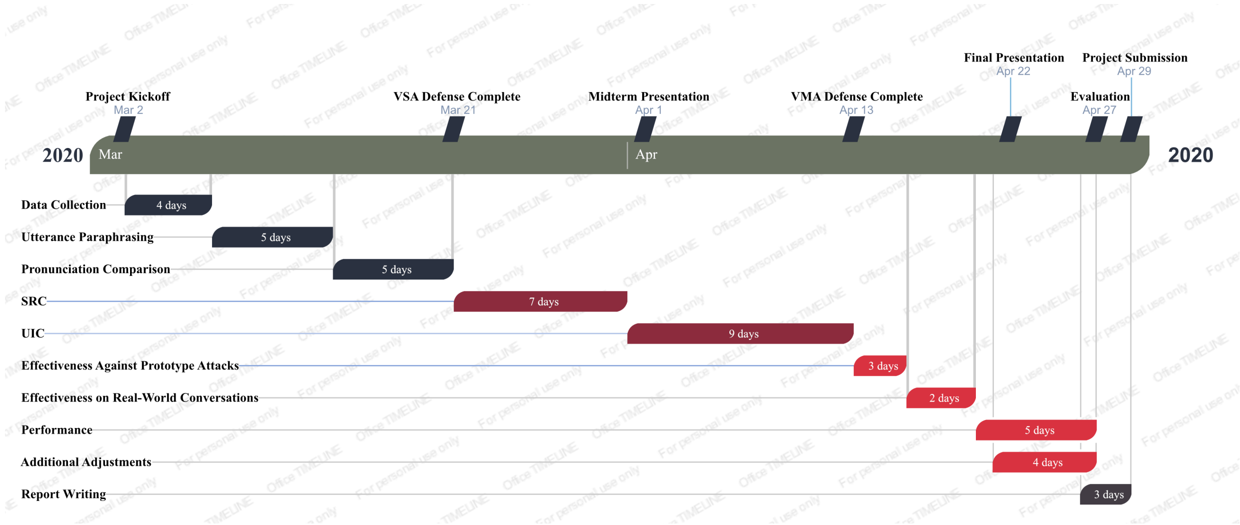

Fig. 1. A tentative schedule of the project development, divided into three four separate segments: defense against VSA attack, defense against VMA attack, evaluation, and reporting observations and conclusions.

## References

[1] D. Kumar, R. Paccagnella, P. Murley, E. Hennenfent, J. Mason, A. Bates, and M. Bailey, “Skill squatting attacks on amazon alexa,” *27th USENIX Security Symposium*.

[2] N. Zhang, X. Mi, X. Feng, X. Wang, Y. Tian, and F. Qian, “Dangerous skills: Understanding and mitigating security risks of voice-controlled third-party functions on virtual personal assistant systems,” *2019 IEEE Symposium on Security and Privacy (SP)*, 2019.